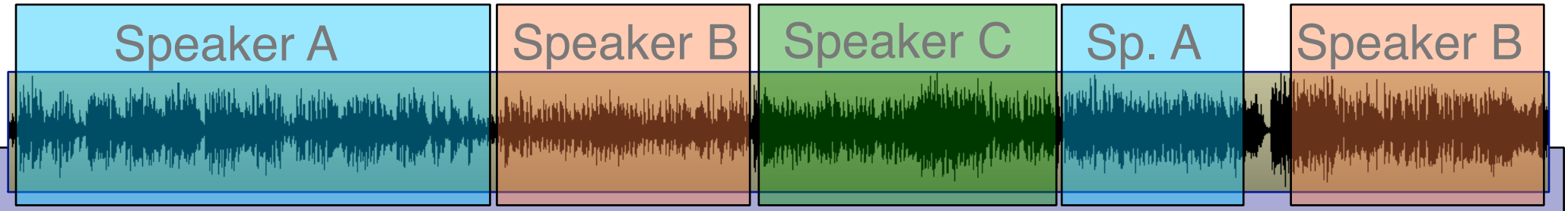


AMI RT06s
SAD and SPKR submission

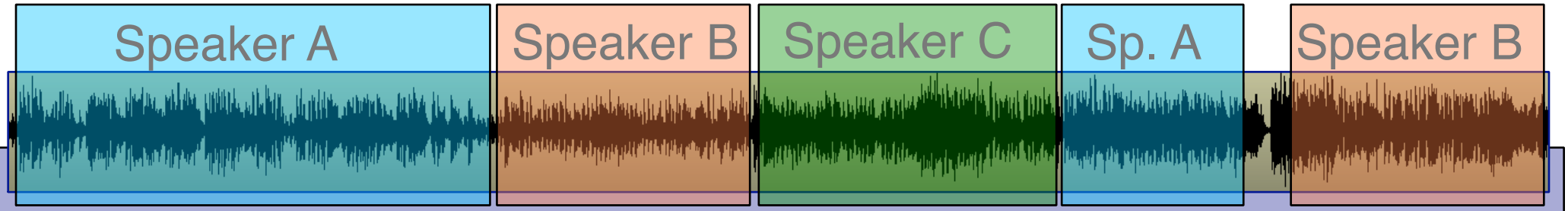
David van Leeuwen
TNO Human Factors

Marijn Huijbregts
University of Twente

AMI
AUGMENTED MULTI-PARTY INTERACTION

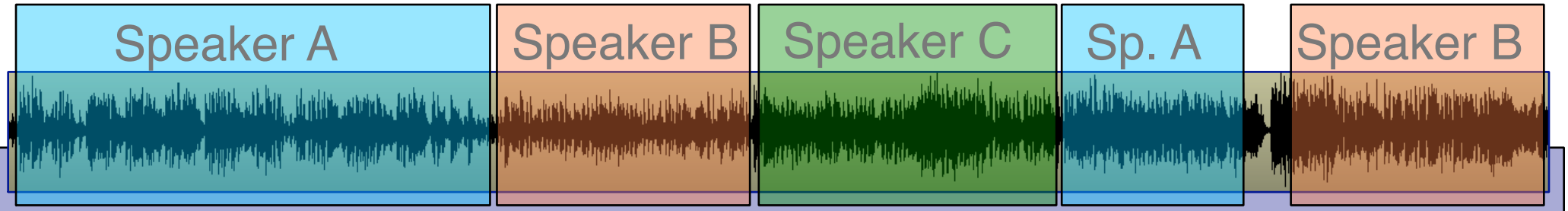
TNO | Knowledge for business

University of Twente
The Netherlands



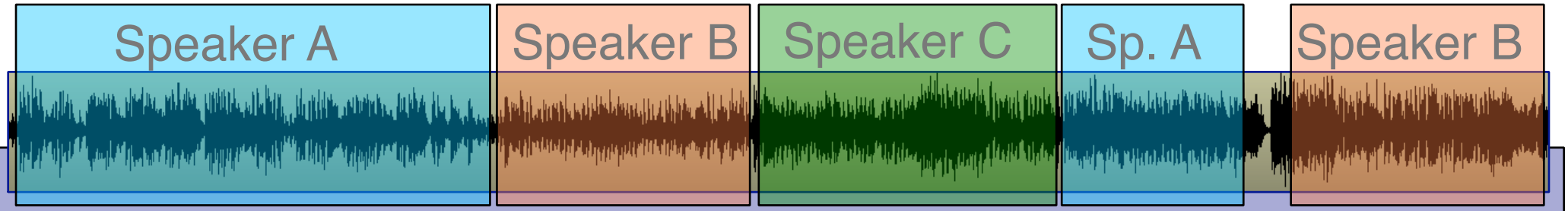
Contents

- System overview
- Databases
- Speech Activity Detection
- Speaker diarization systems
  - BIC + resegmentation (Primary)
    - Multiple speaker outputs (not submitted)
  - HMM+BIC
  - Cut & mix
    - Some analysis



System overview

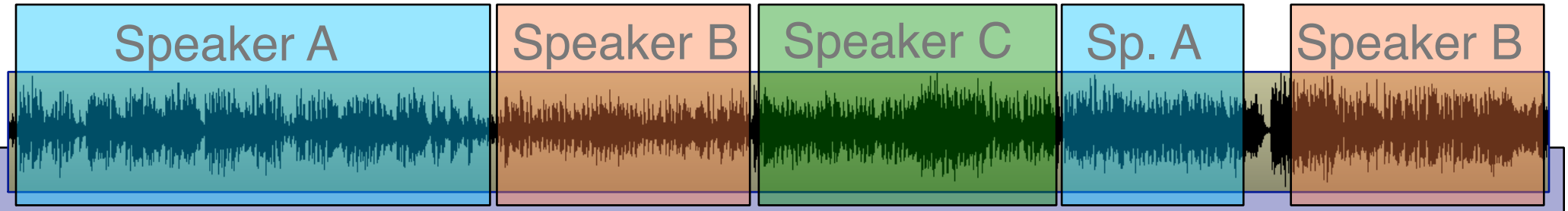
- Microphone signal from either
  - Optional Microphone Array Beamforming (Mike Lincoln)
  - One central microphone (the one from SDM)
- Speech activity detection (SAD)
- SPKR system either
  - segmentation → clustering → resegmenation
  - iterative
    - resegmentation
    - cluster reduction



Databases

(more important than one might think)

- Training:
  - Only for SAD
  - 10 AMI meetings (RT05s dev test)
    - not any other meetings (RT04s, RT05s)
    - slightly detrimental to SAD performace
- Development test
  - RT05s meeting room data
    - none of the Lecture room data

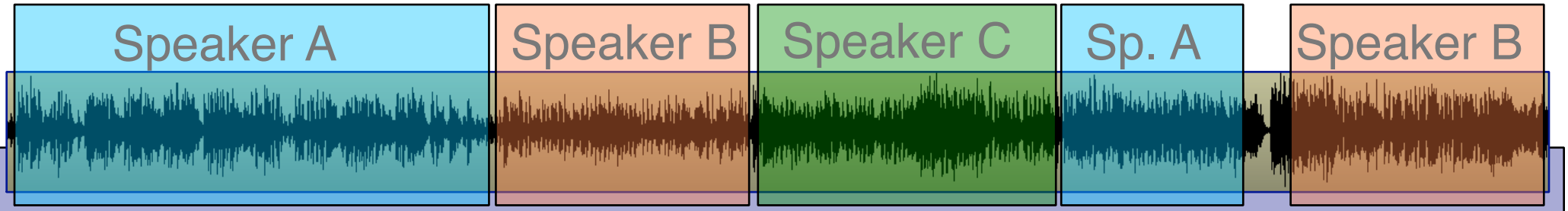


The SAD story

- Two-state HMM decoder
  - One for silence (non-speech)
  - one for speech
- 16 mixture GMM output probabilities
- single training of GMMs (no resegmentation or Baum-Welch training)
- fixed intra-state transition probabilities
  - skewed towards speech (1:10)
- Trained on 10 AMI RT05s dev-test meetings
- Enormous bug from RT05s system removed

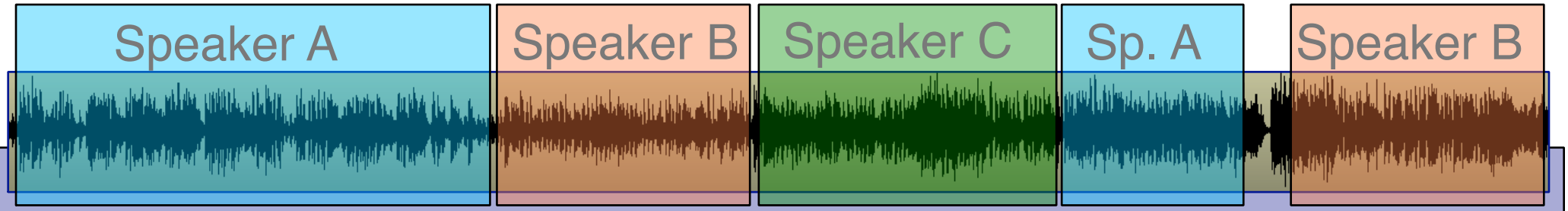
MDM ≡ SDM

conf	lect
4.3%	22.8%



A really sad story

- The history of the TNO meeting in RT06s...

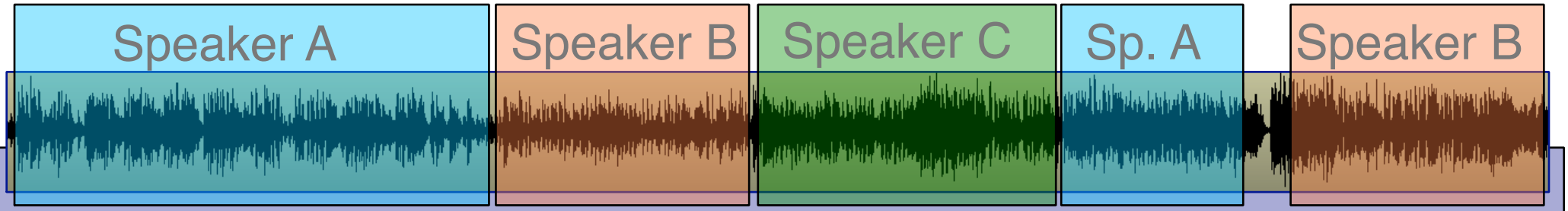


Speaker diarization (primary)

- Tried many variants
  - ICSI RT05s
  - LIMSI / Cambridge Broadcast News clustering
- Reverted to TNO RT05s system
  - BIC segmentation, λ=1.6
  - BIC clustering, λ=6
- followed by
  - Viterbi resegmentation
  - including silence from SAD

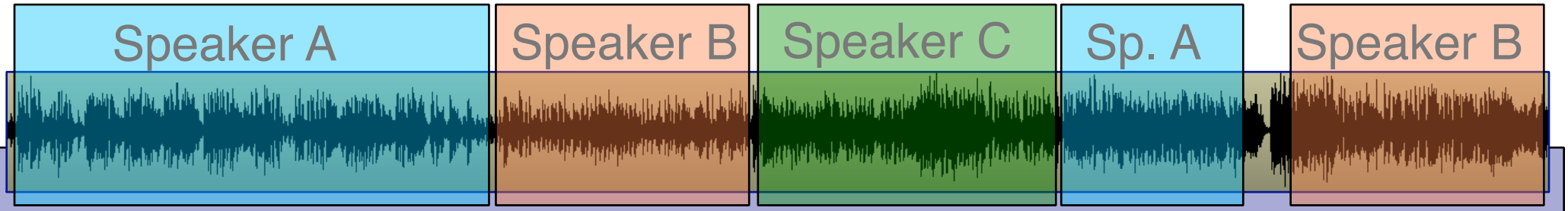
MDM ≡ SDM

conf		lect	
overl	no	overl	no
44.8	32.6	27.8	27.4



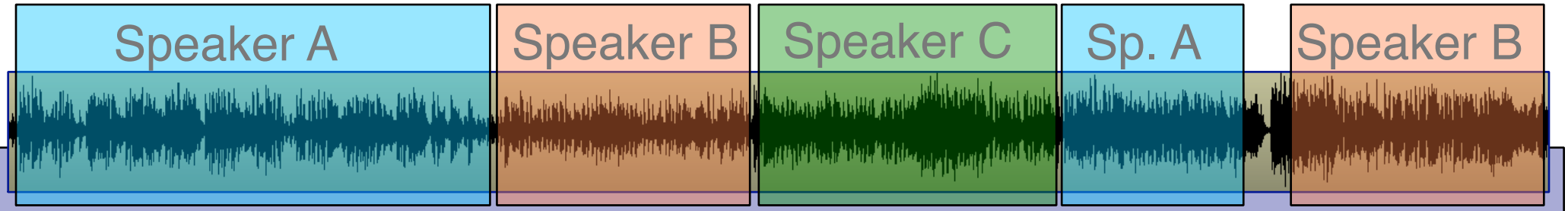
Simultaneous speaker output

- One state/cluster + Viterbi
  - max. one simultaneous speaker hypothesis
- Approach 1
  - generate $\binom{N}{2}$ speaker-combo states
  - allow transitions from single ↔ combo including single
  - re-segment
- Approach 2
  - Guess t seconds overlap
- Not working/submitted
  - features are not linear

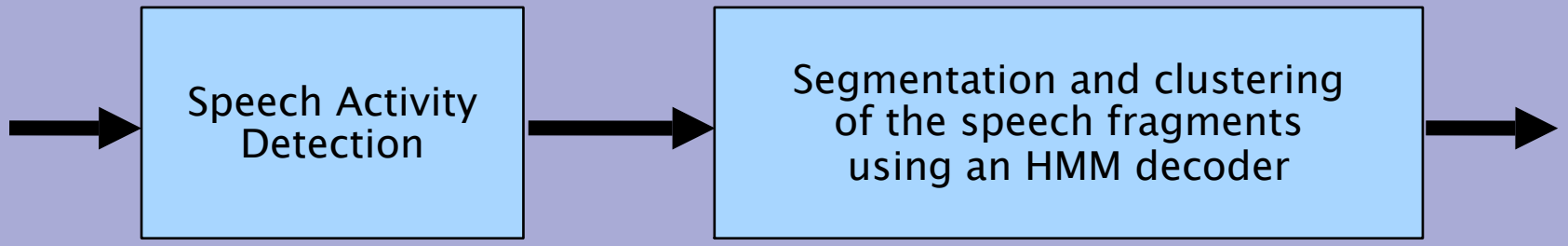


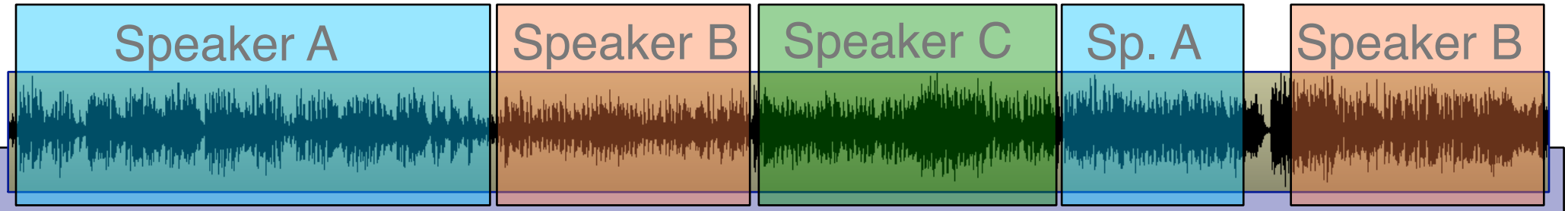
Two contrastive systems

- Based on HMM/BIC (hmm_bic)
- Based solely on HMM(cut&mix)

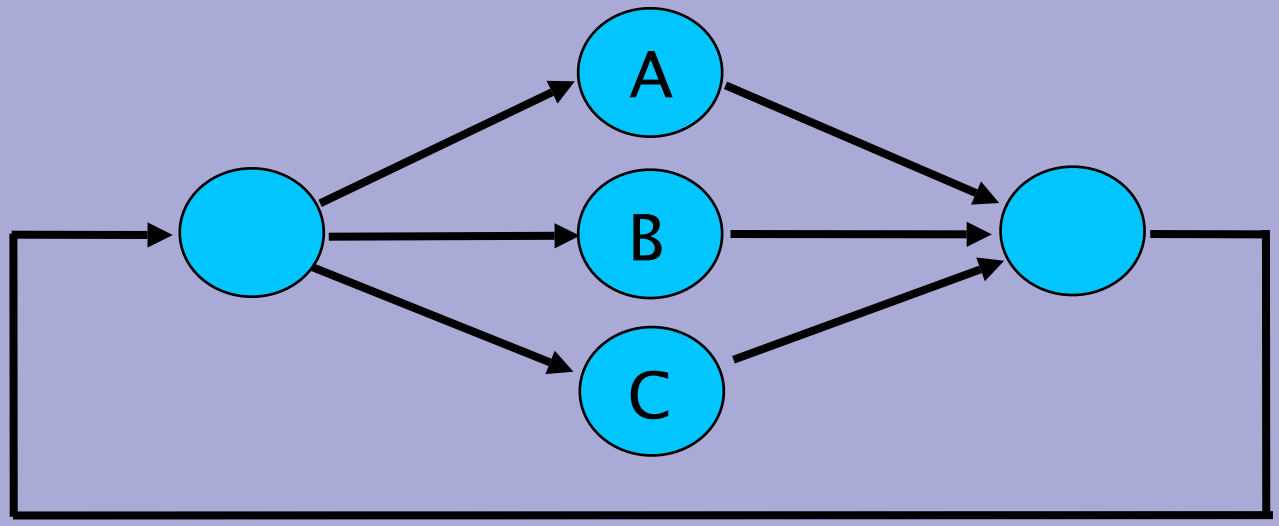


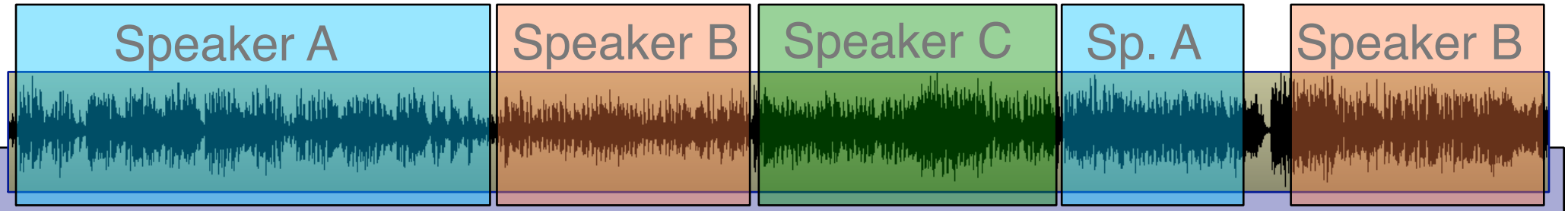
The HMM-BIC system



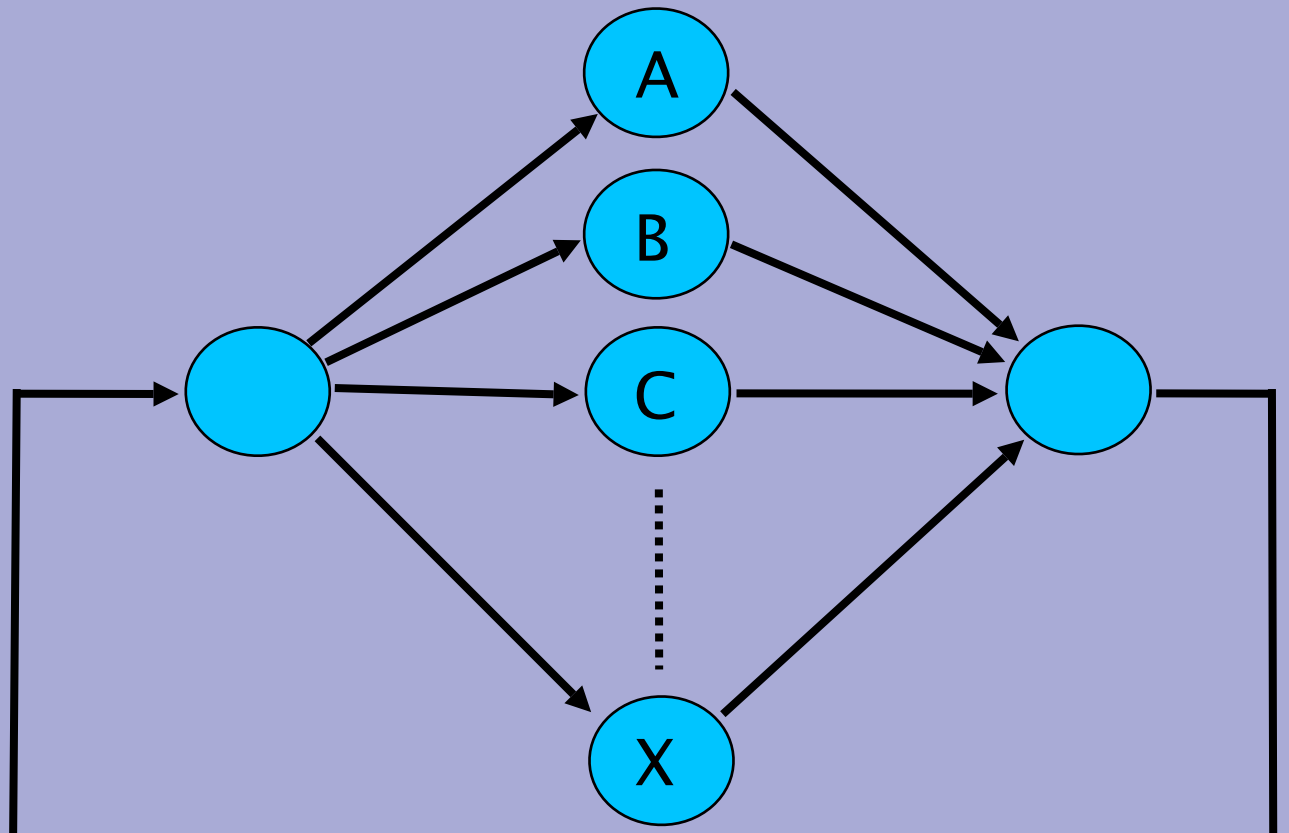


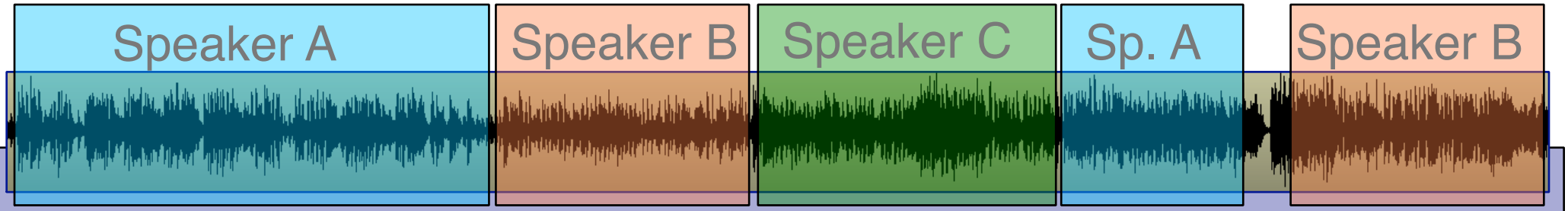
The (trained) HMM architecture



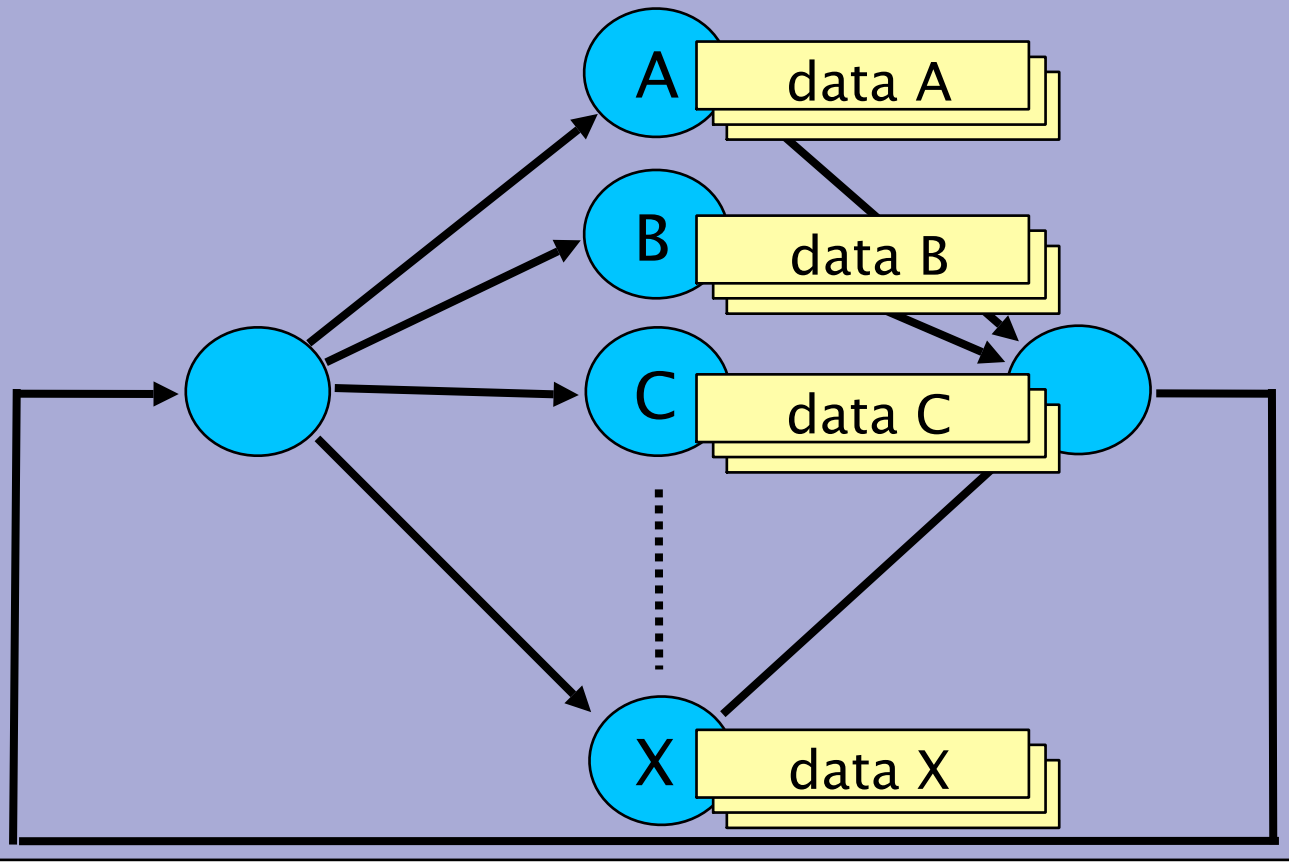


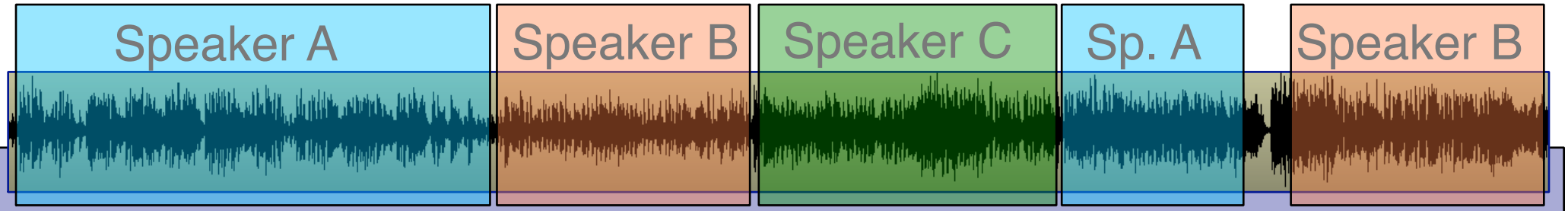
The initial HMM



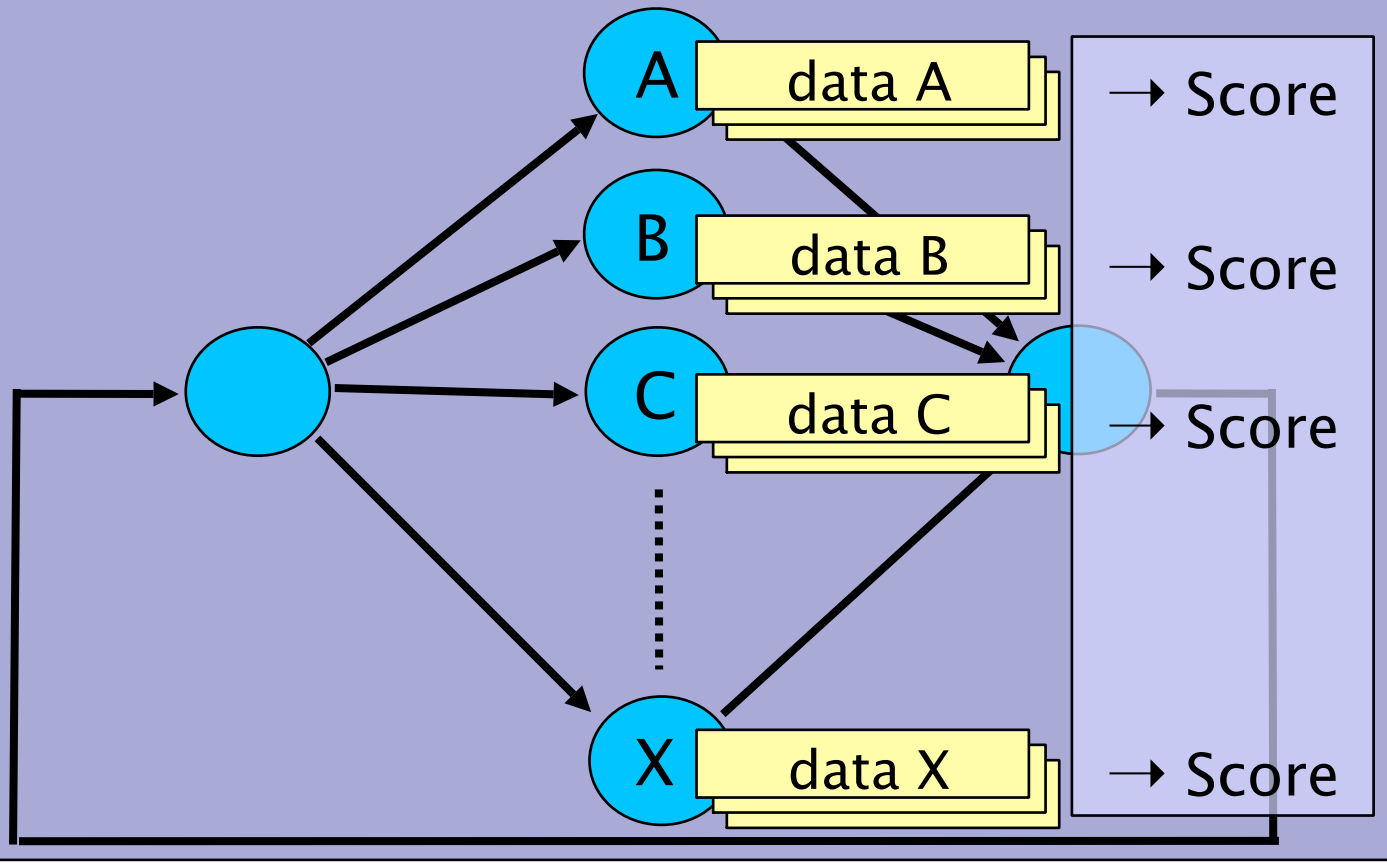


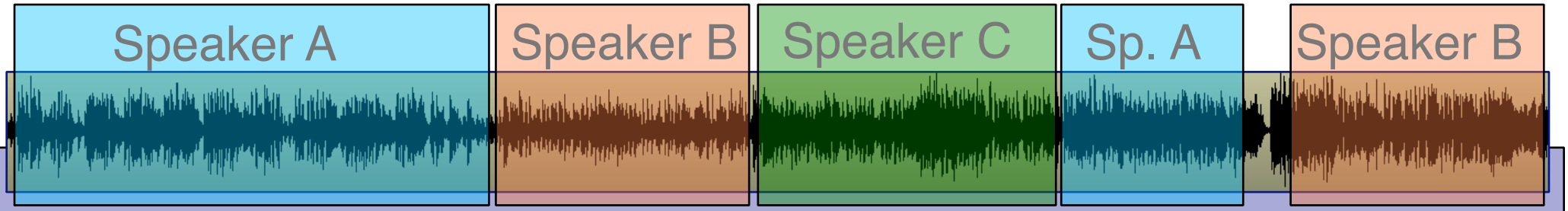
Step 1: Align data and train



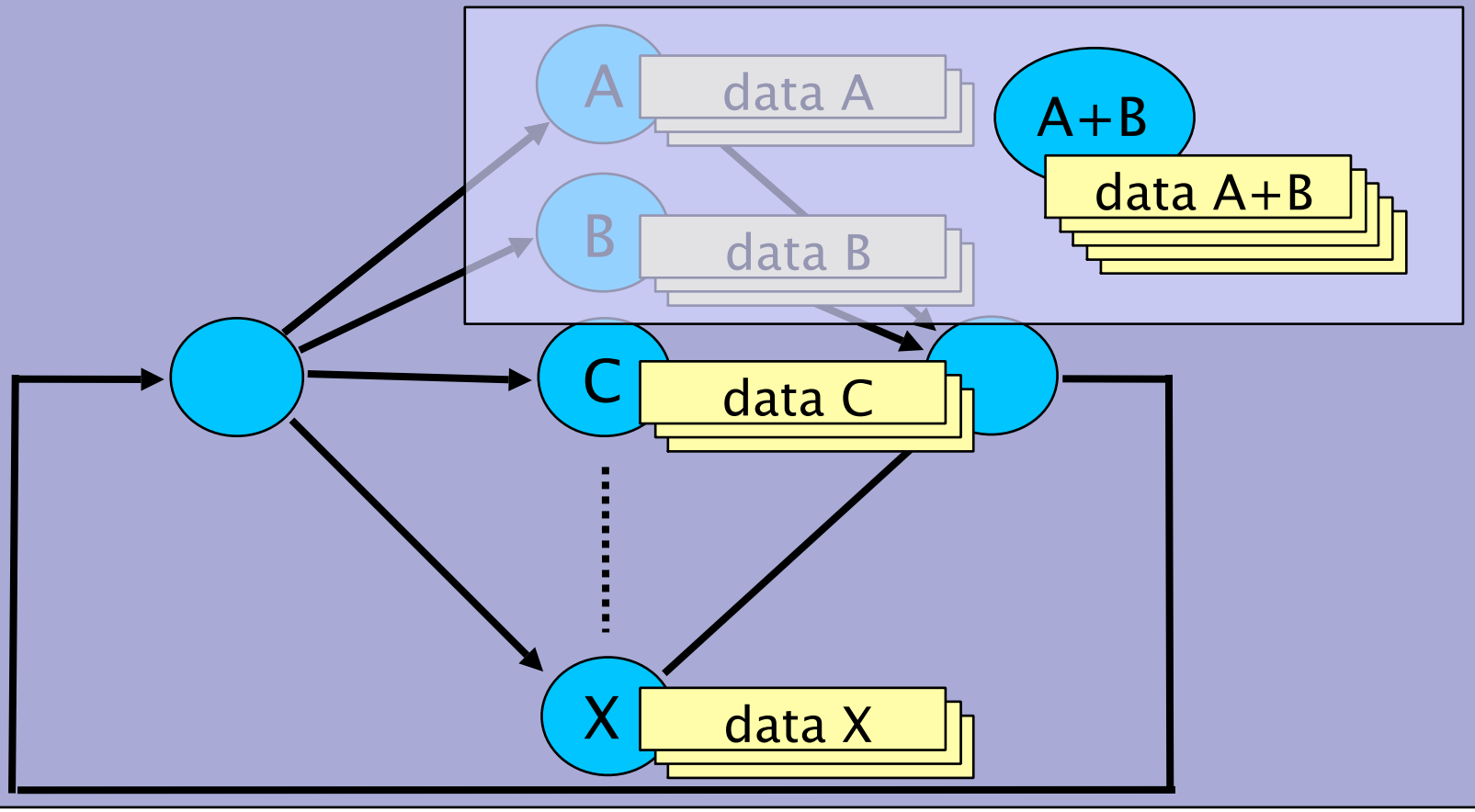


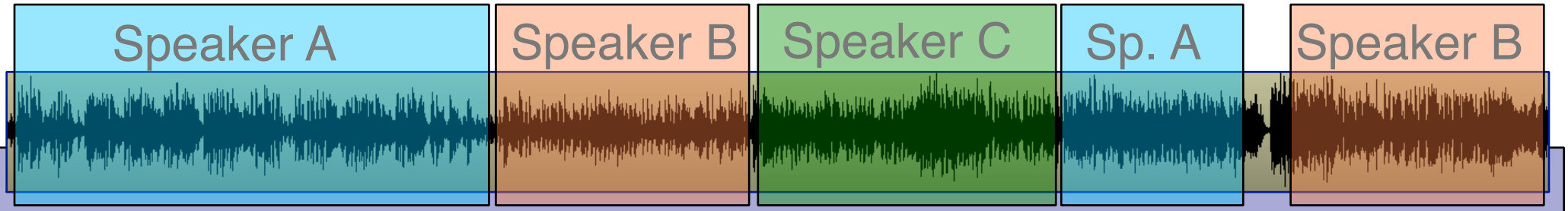
Step 1: Align data and train



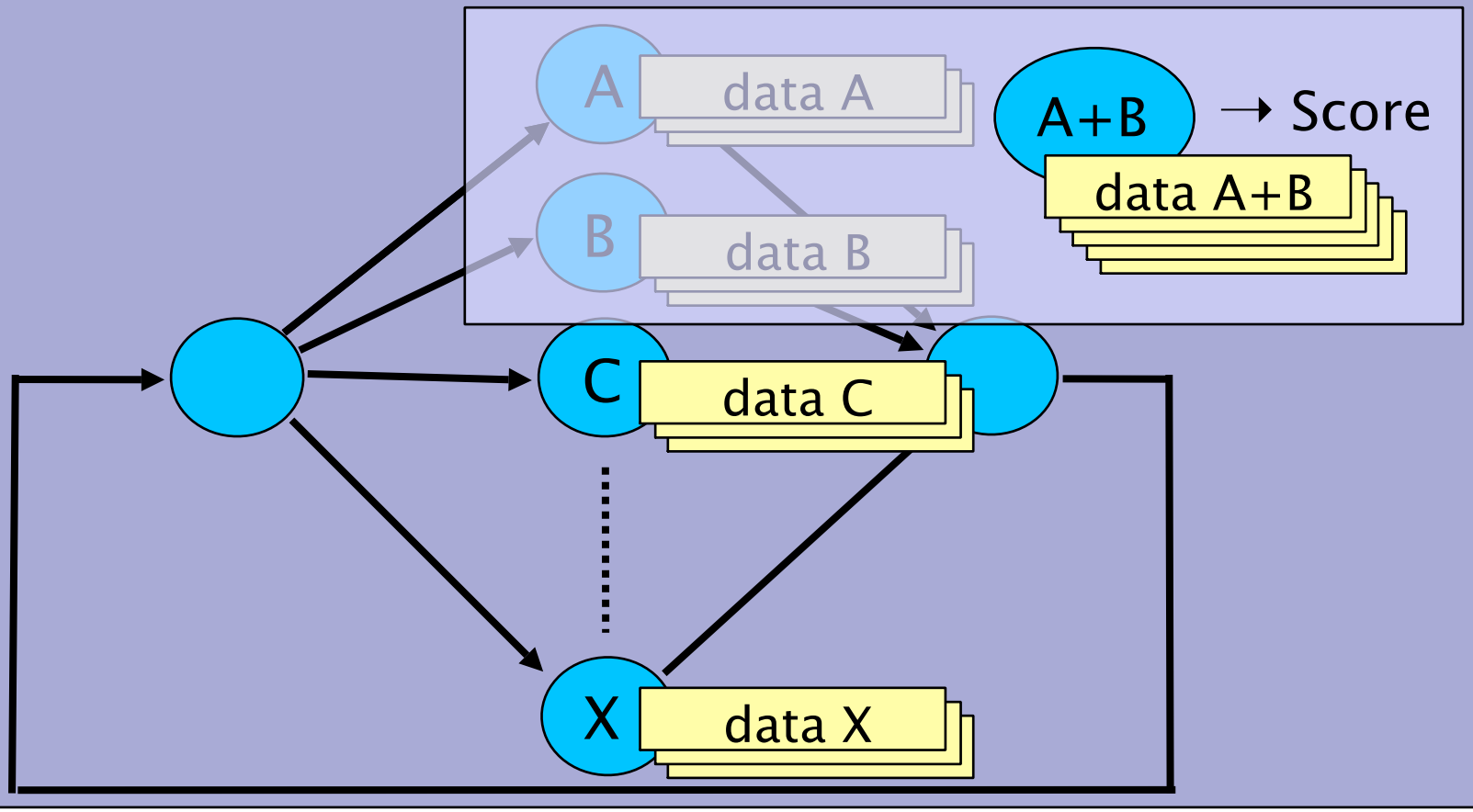


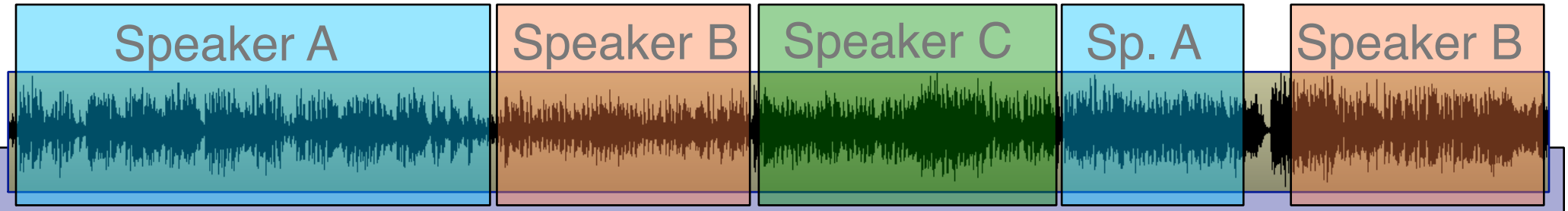
Step 2: Merging





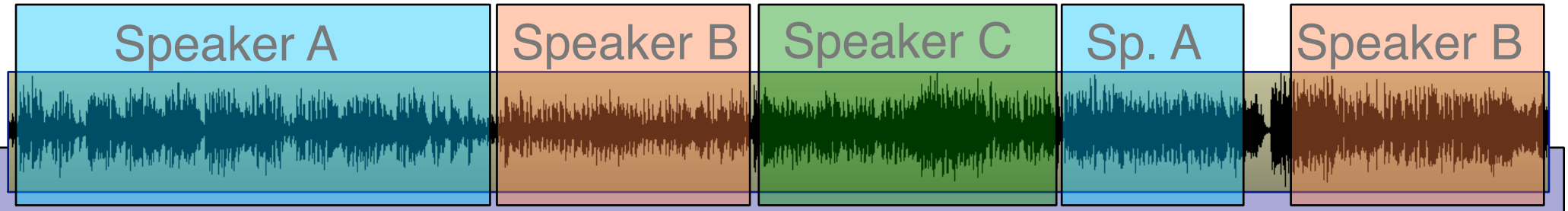
Step 2: Merging





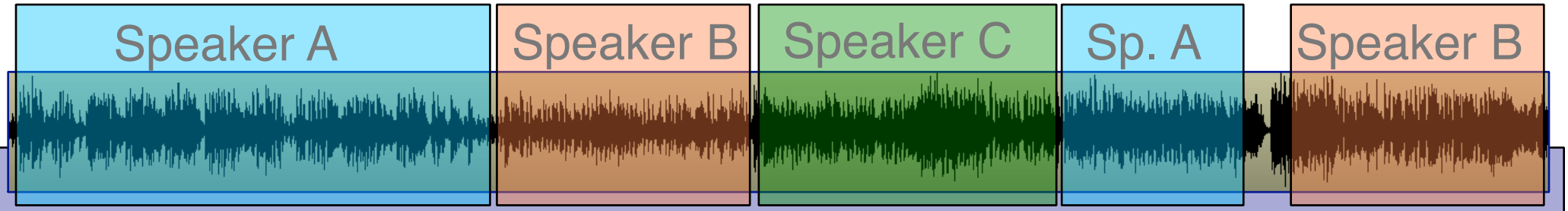
Step 3: Determine maximum BIC score

- For each combination of models calculate BIC
  BIC = Score(AB) – Score(A) – Score(B)
- Stop merging if no BIC score is bigger than zero
- Otherwise merge the models with the biggest BIC and start a new merging iteration

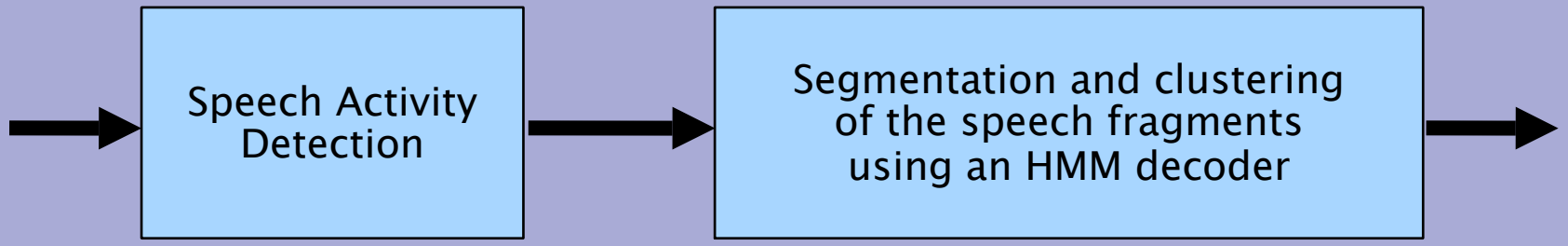


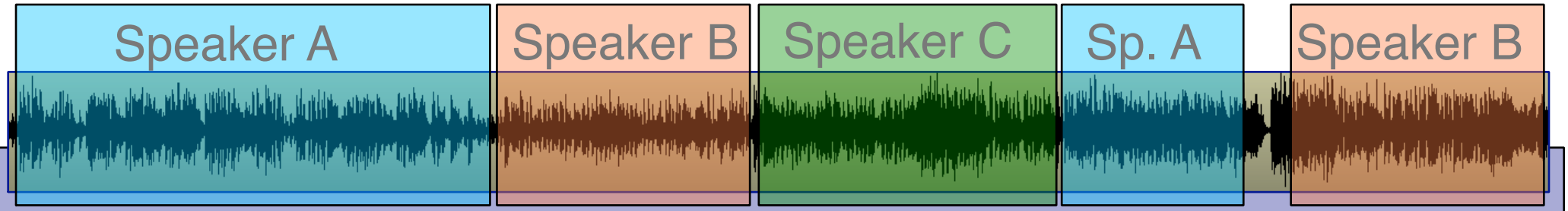
The HMM-BIC system

- 4.63 times real-time on a 2.8GHz Intel Xeon processor
- The merging step takes most processor time
- It is not possible to divide data over multiple states

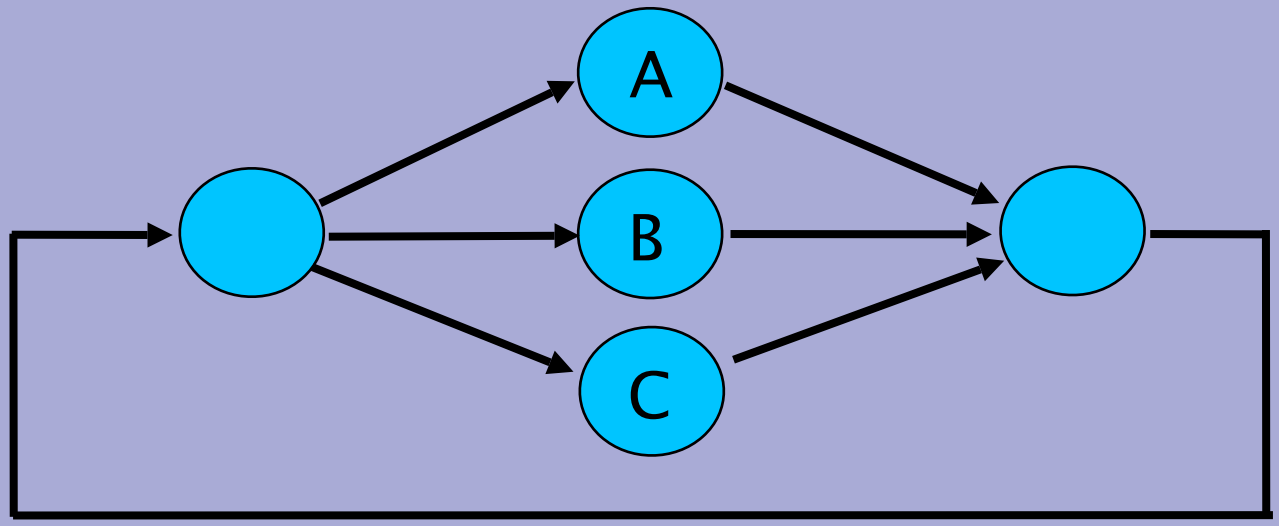


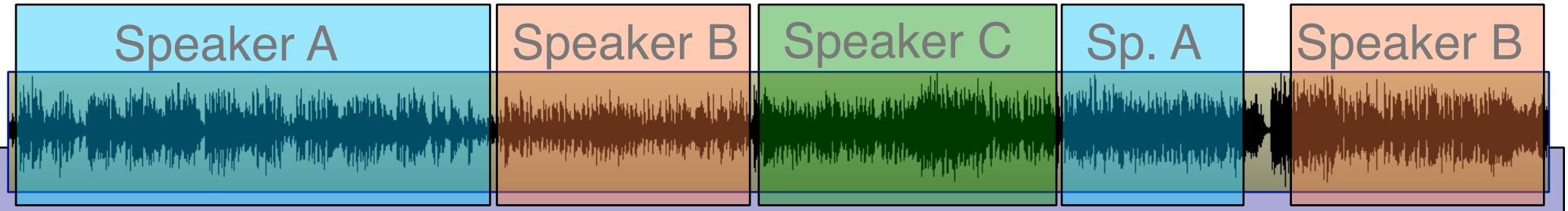
The Cut&Mix system



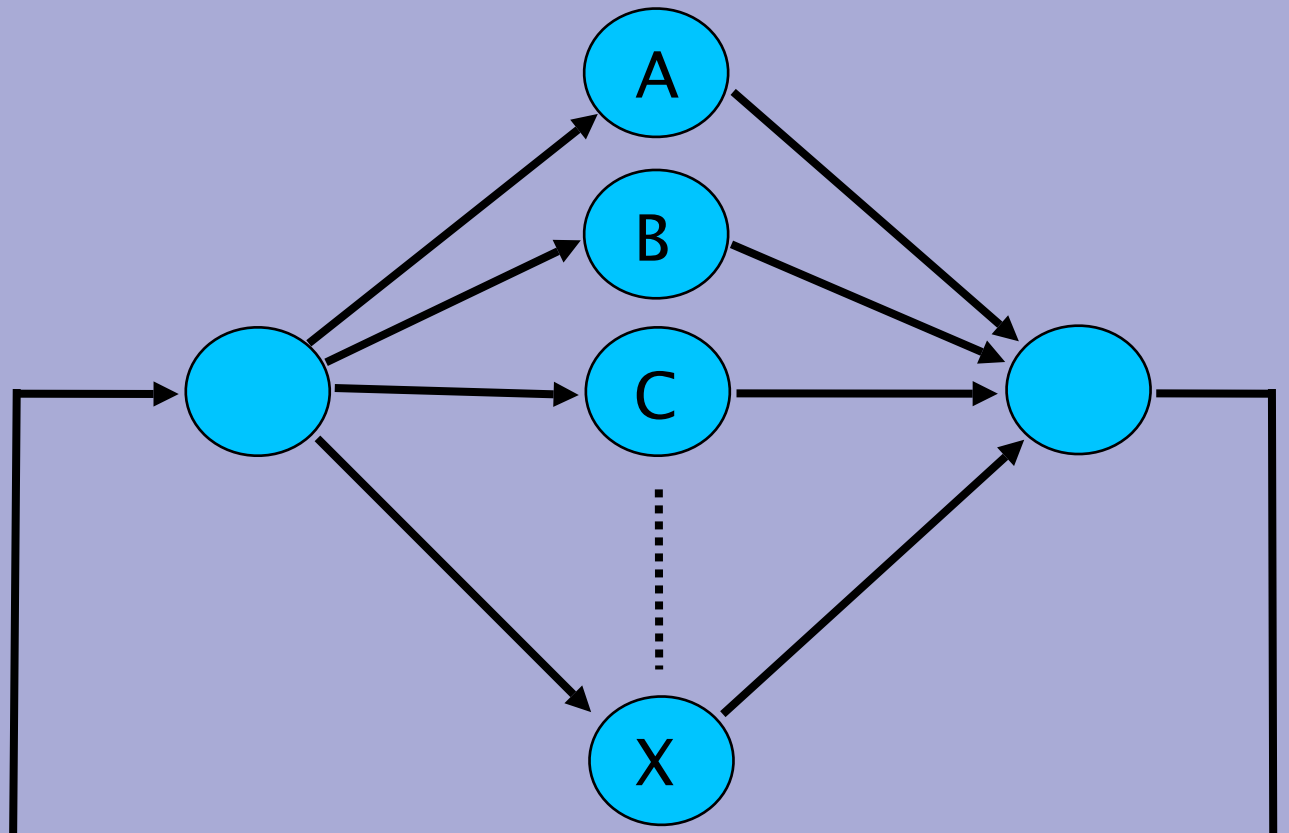


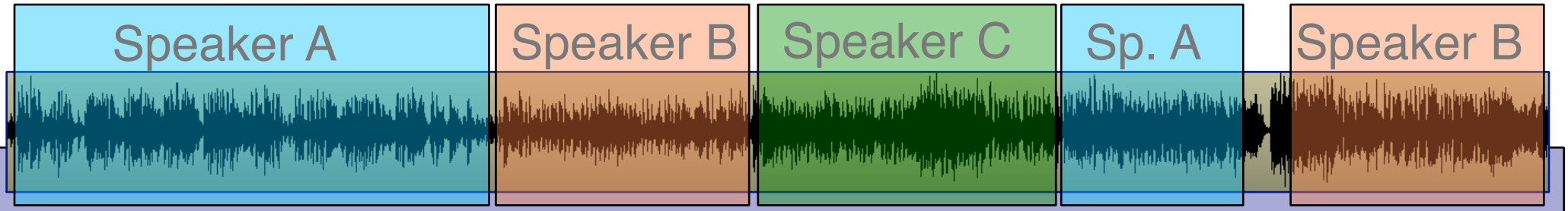
The (trained) HMM architecture



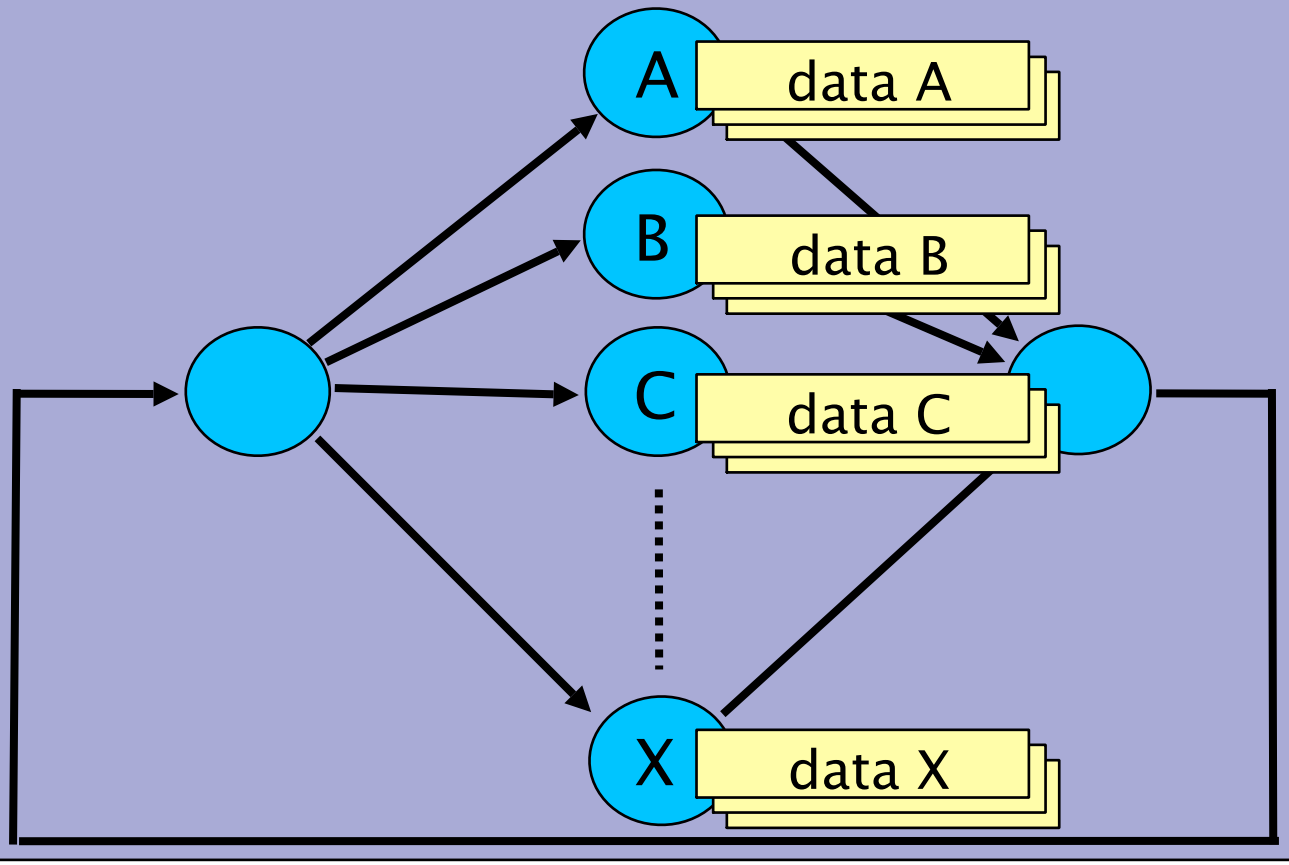


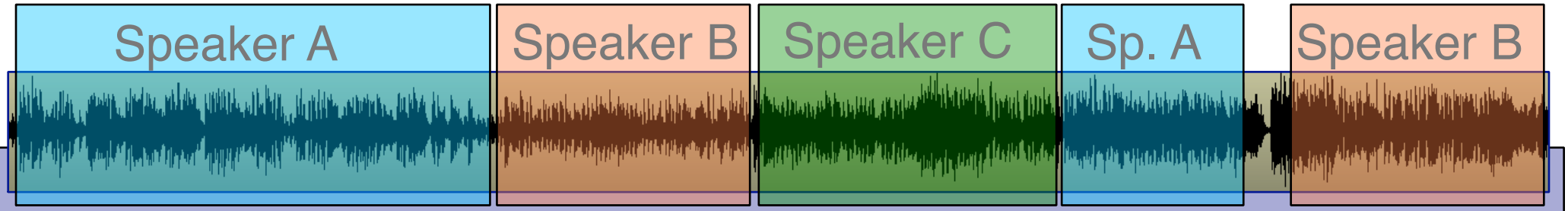
The initial HMM



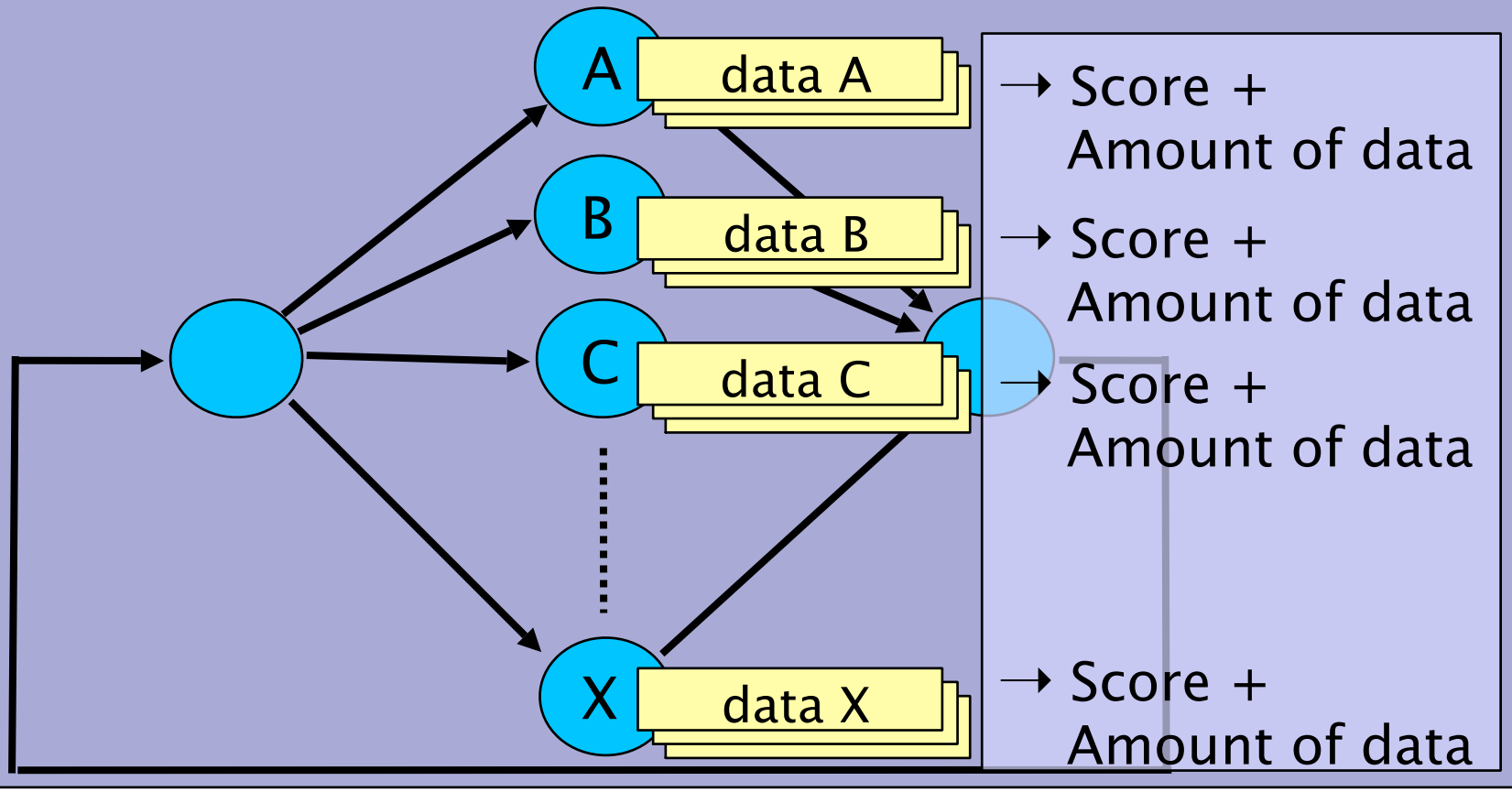


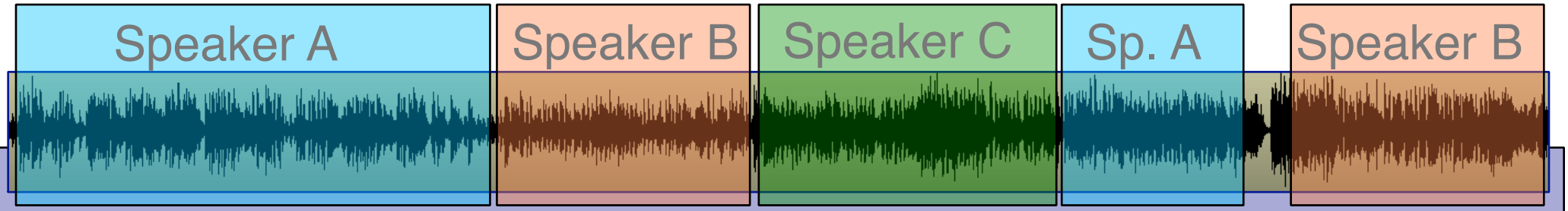
Step 1: Align data and train



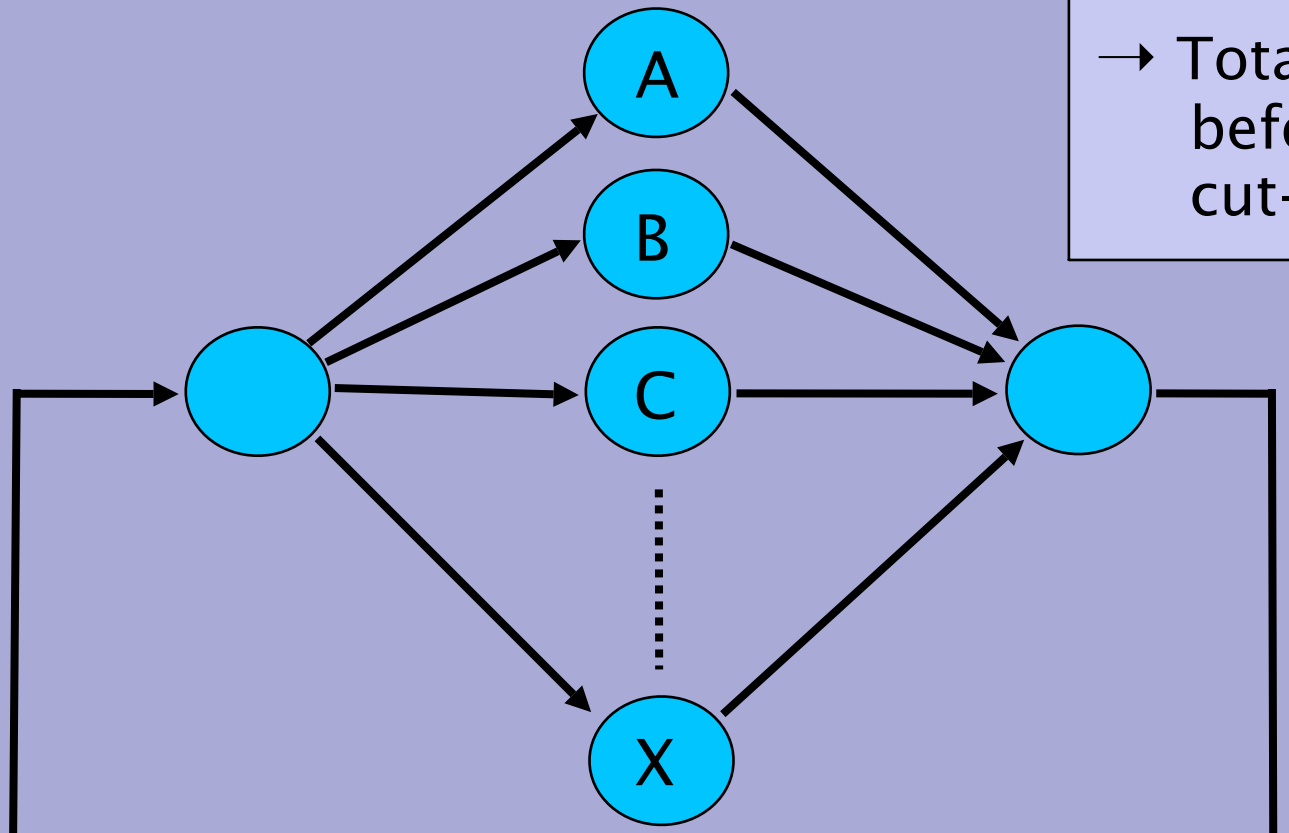


Step 1: Align data and train

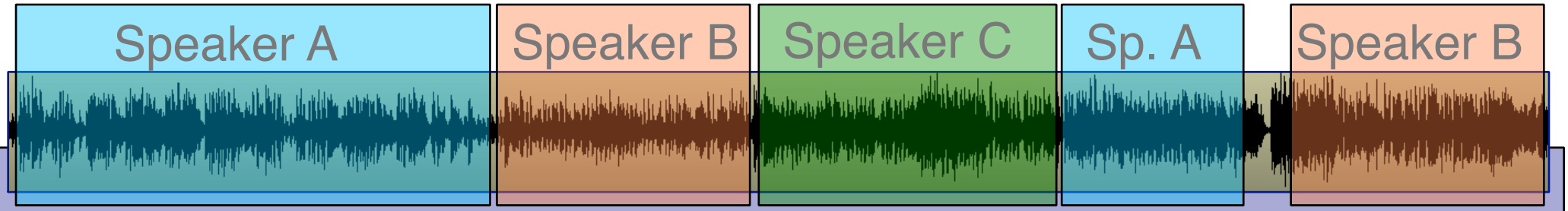




Step 2: Cut

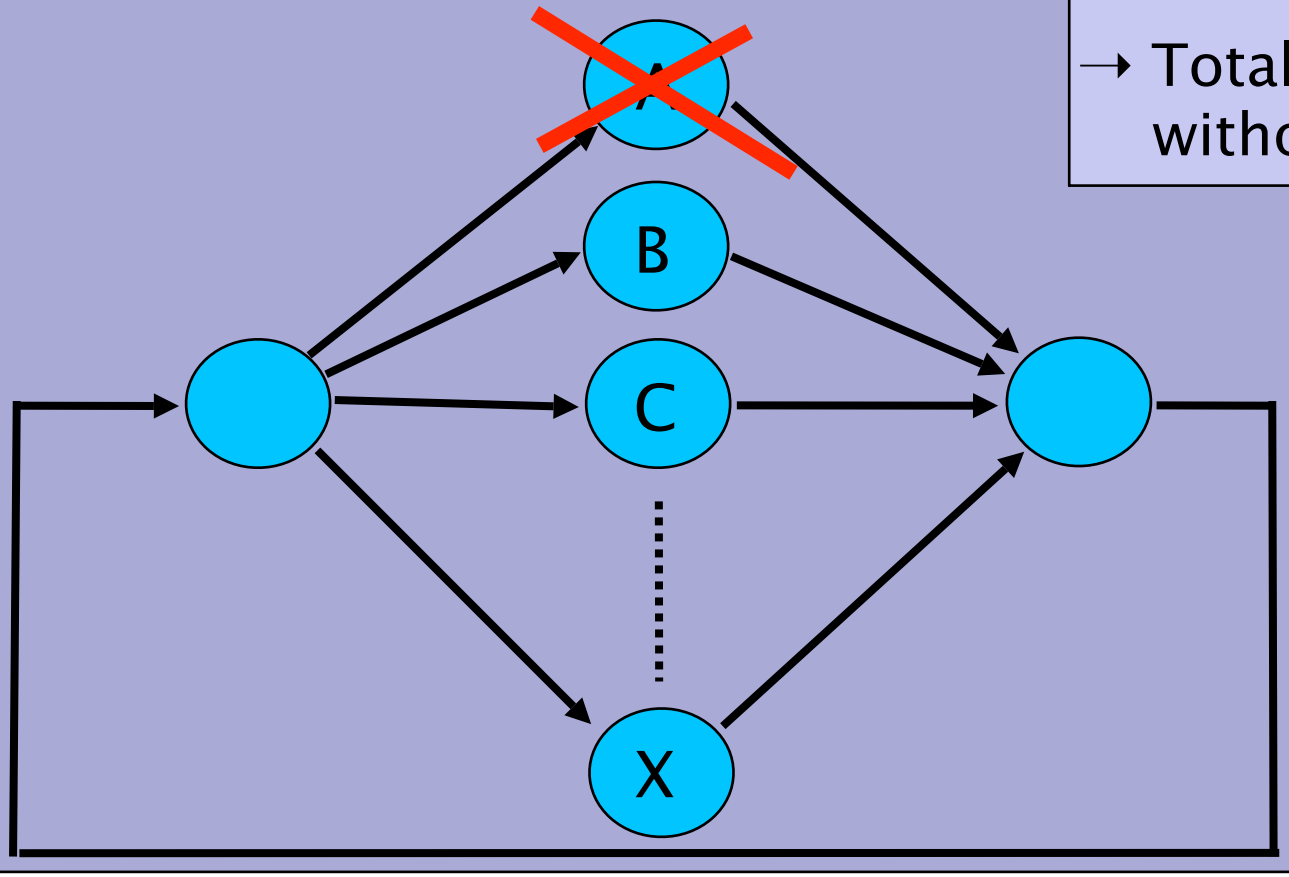


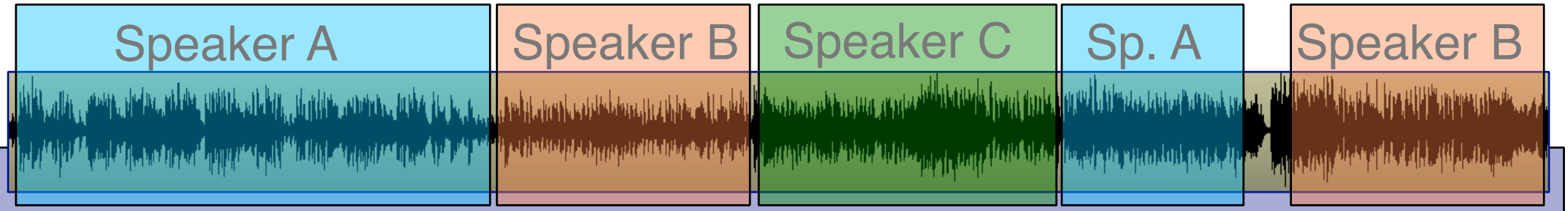
→ Total score before cut-procedure



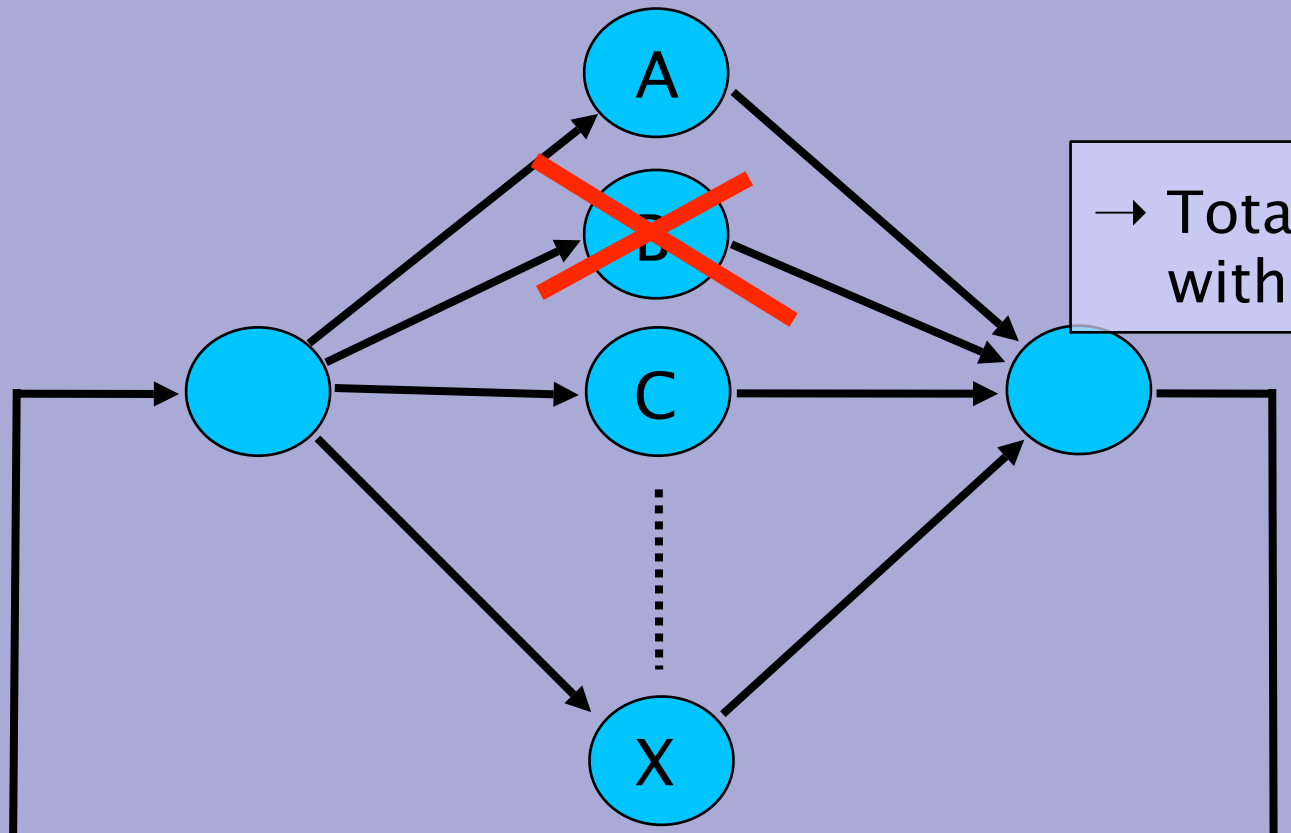
Step 2: Cut

→ Total score without state A

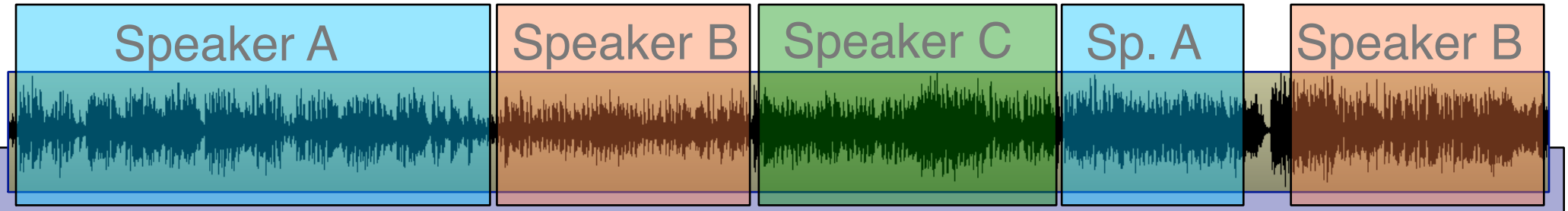




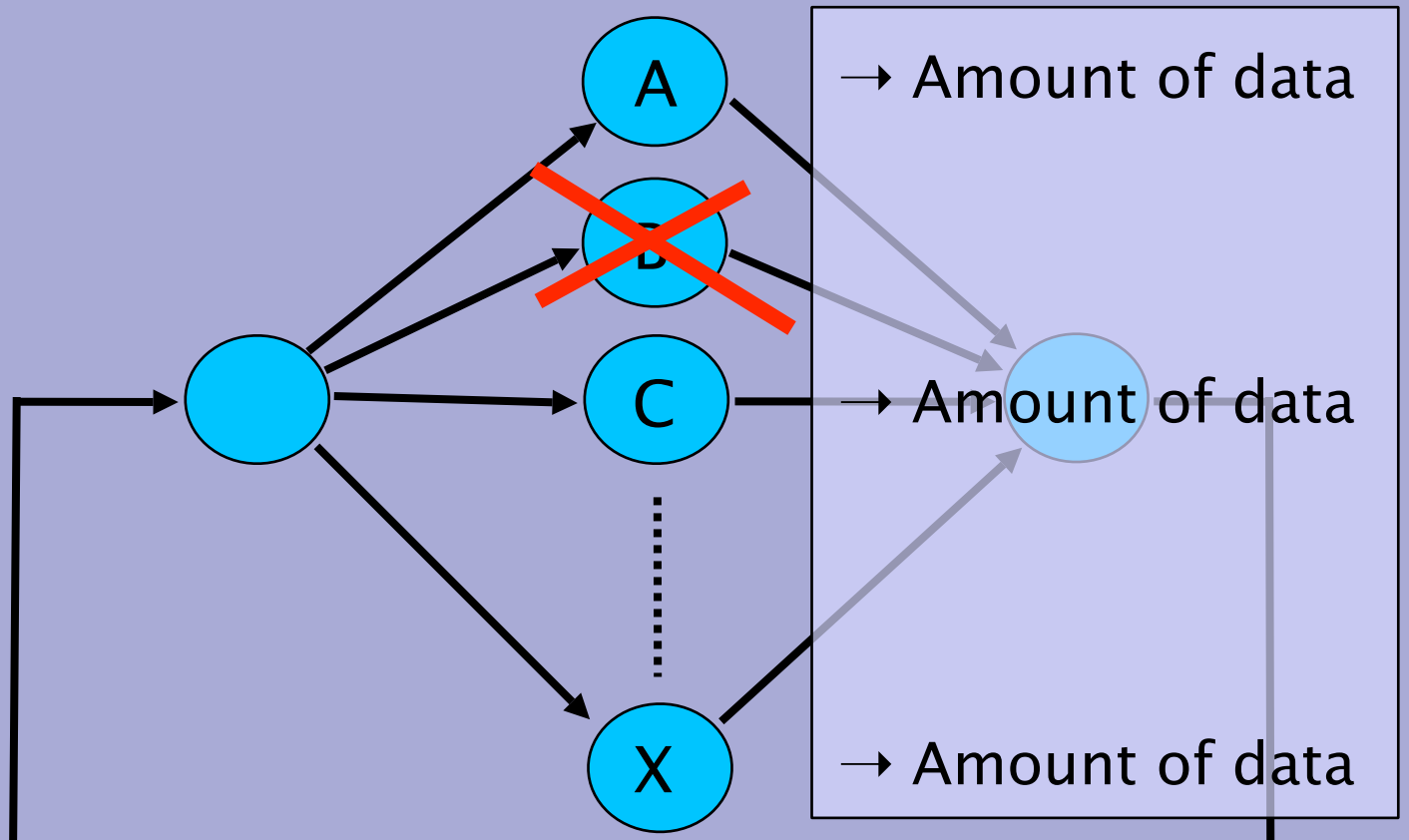
Step 2: Cut

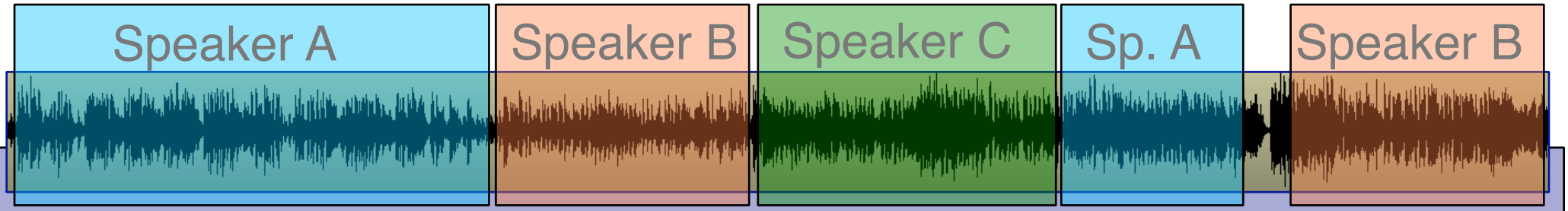


→ Total score without state B


Speaker A
Speaker B
Speaker C
Sp. A
Speaker B

Step 2: Cut

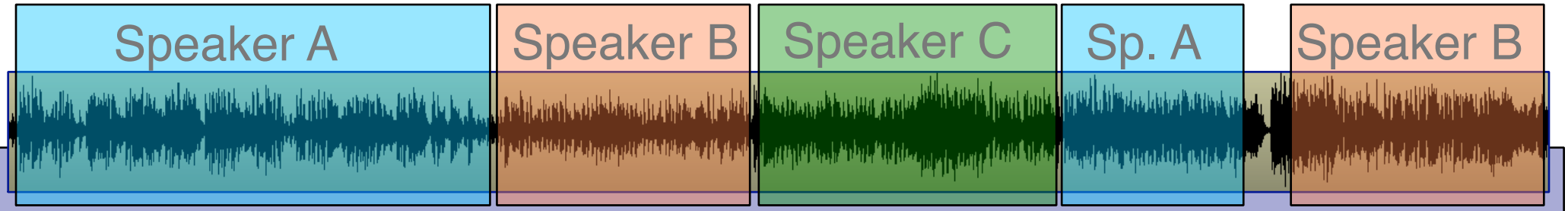

A
B
C
X
→ Amount of data
→ Amount of data
→ Amount of data



Step 3: Mix

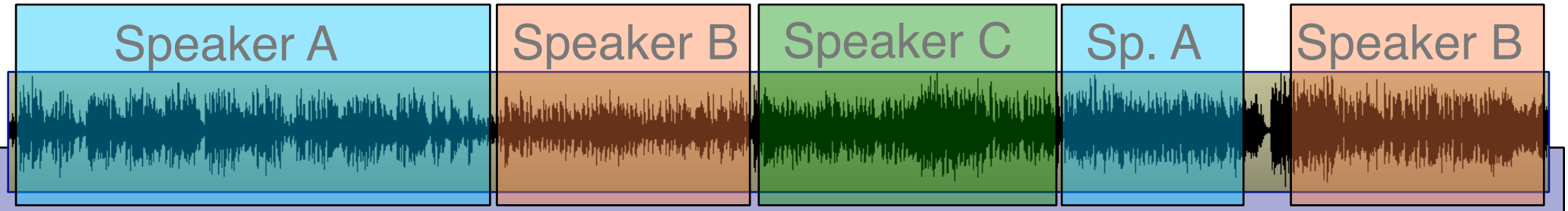
- The best performing system from the previous step serves as input for the 'mix-step' (state X cut away)
- The total number of gaussians in the system should stay the same
- The gaussians from state X will be distributed between the remaining states using the formula:

$$\#extraGaussians(A) = (\#dataAfter(A) - \#dataBefore(A)) \; / \; (\#dataBefore(X) \; / \; \#gaussians(X))$$



Step 3: Mix

- Train all states that received extra gaussians
- Determine the new overall score
- If the new score is better than the original score, start a new cut iteration
- Otherwise, fall back on the previous system (don't cut away state X) and stop



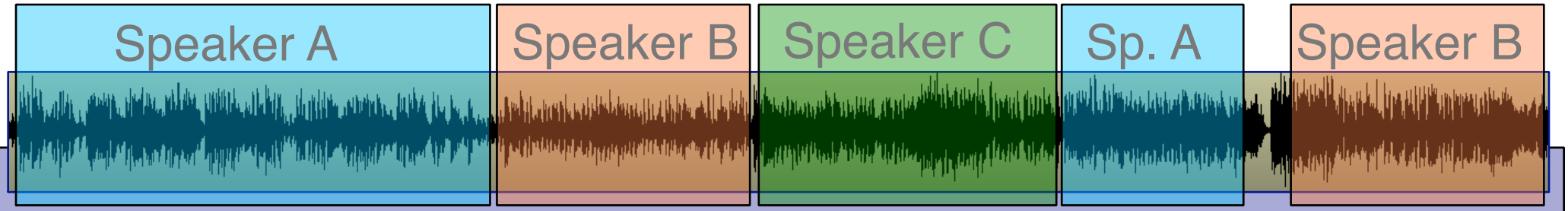
Conference meeting results

note, our MDM ≡ SDM	HMM-BIC	Cut&Mix
Speaker Diarization Error RT06s (overlap)	37.32	39.49
Speaker Diarization Error RT06s (no overlap)	22.90	25.36
Speaker Diarization Error RT05s (no overlap)	21.56	18.60
Real time factor on a 2.8GHz intel Xeon	4.63	2.25

AMI AUGMENTED MULTI-PARTY INTERACTION

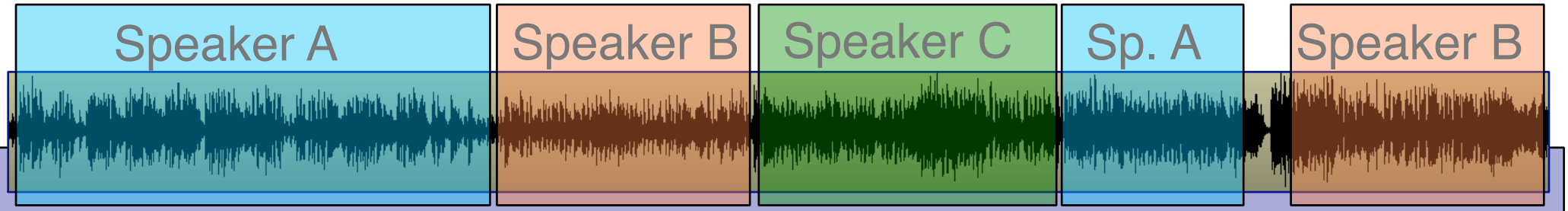
TNO | Knowledge for business

University of Twente
The Netherlands

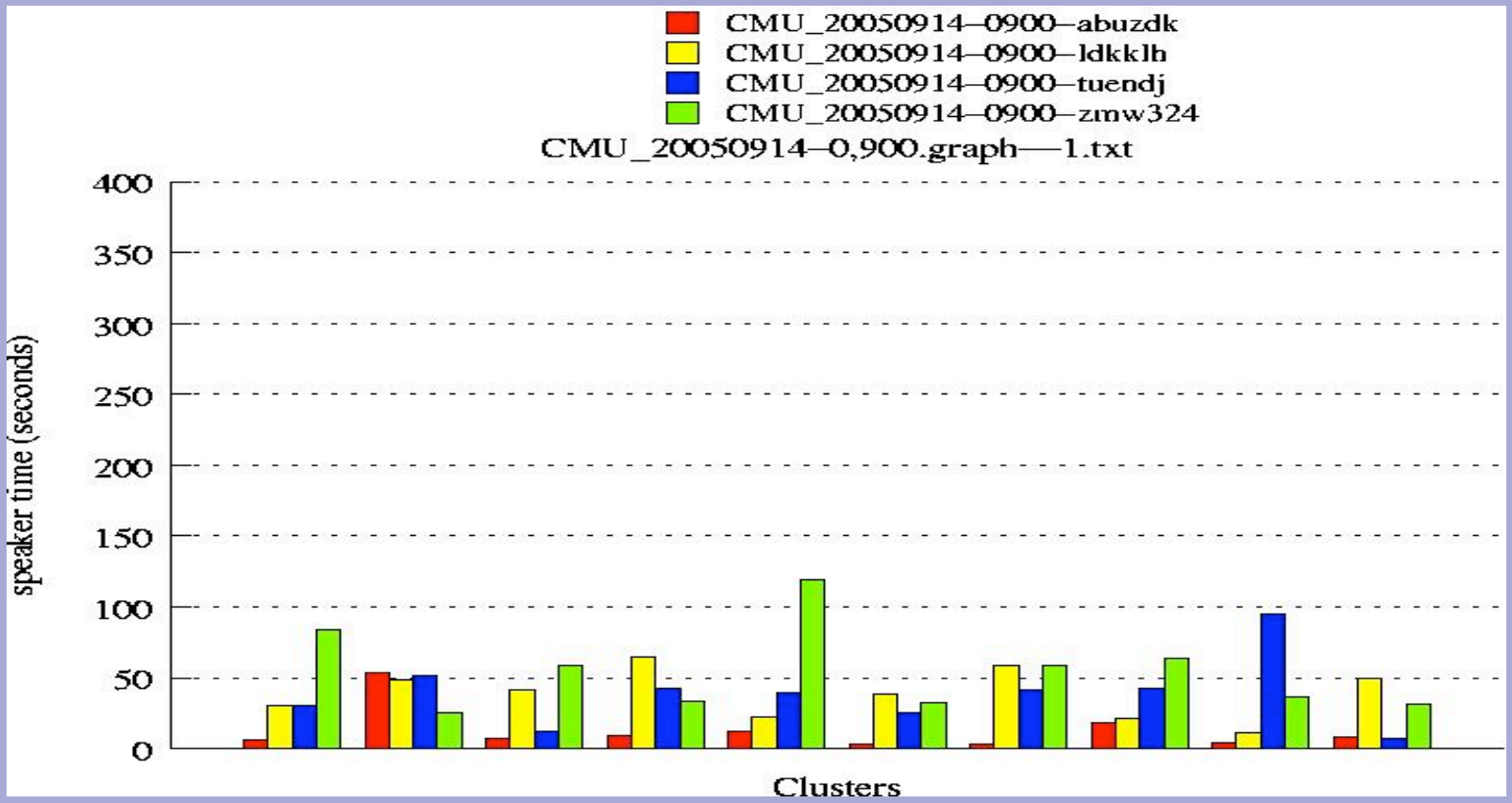


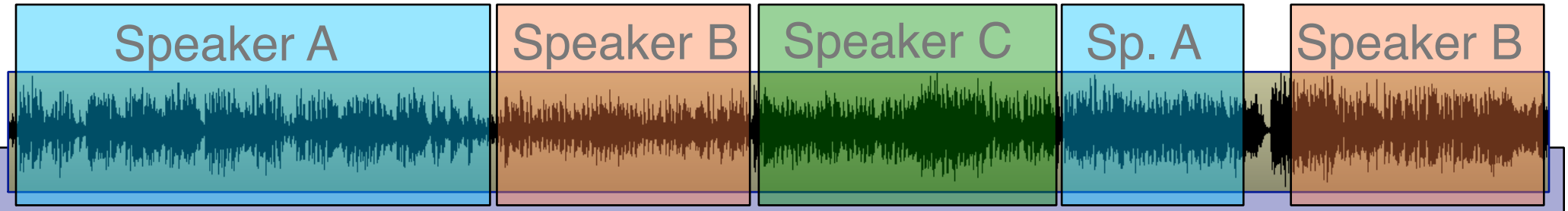
Stop criterion Cut&Mix

- HMM-BIC calculates all possible systems with one state less during merging
- Cut&Mix determines the state that is most likely the best to loose, but it does not calculate each possible next system.
- New stop criterion might help; stop when all possible systems fail...

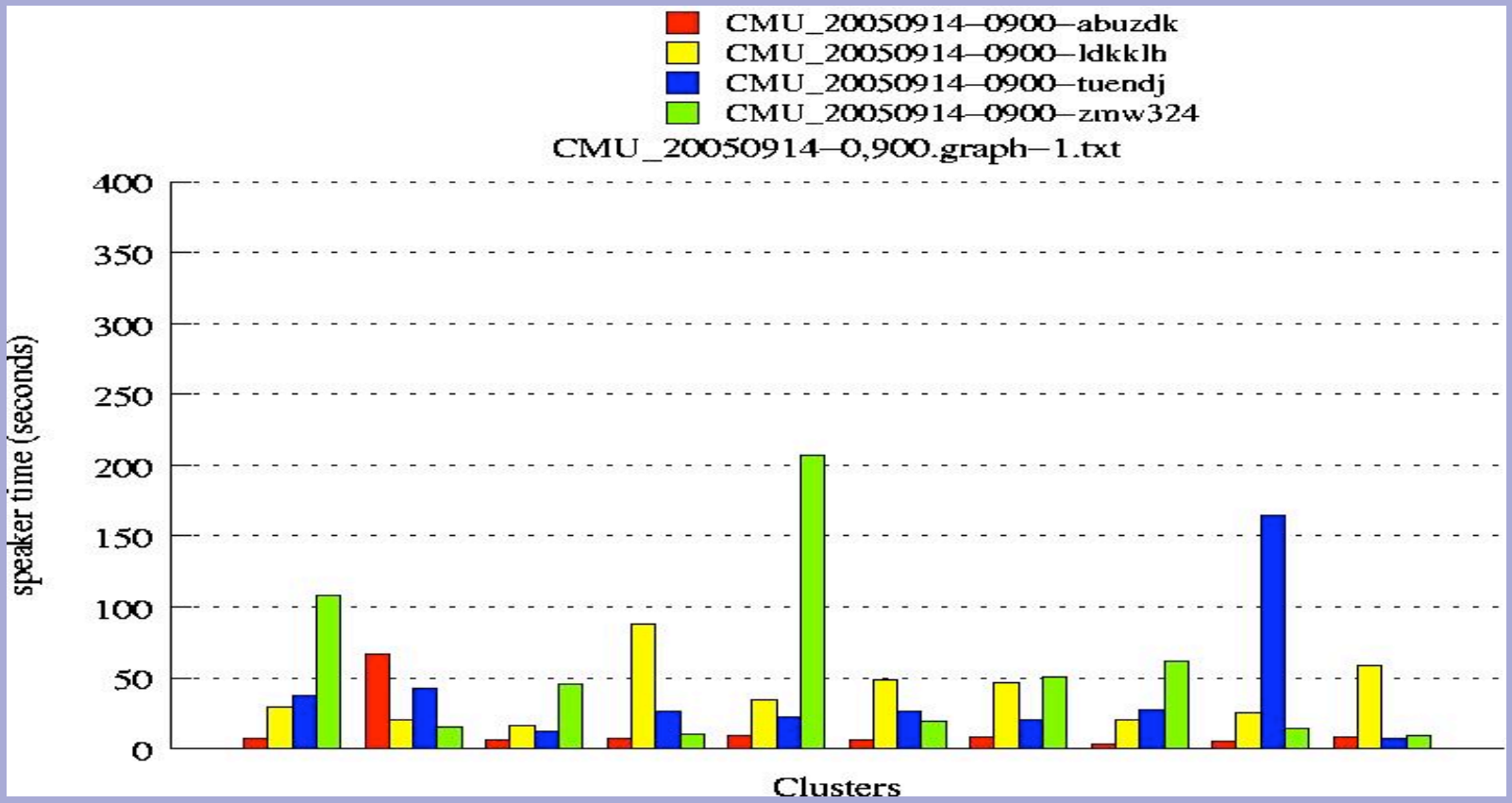


Speaker state domination





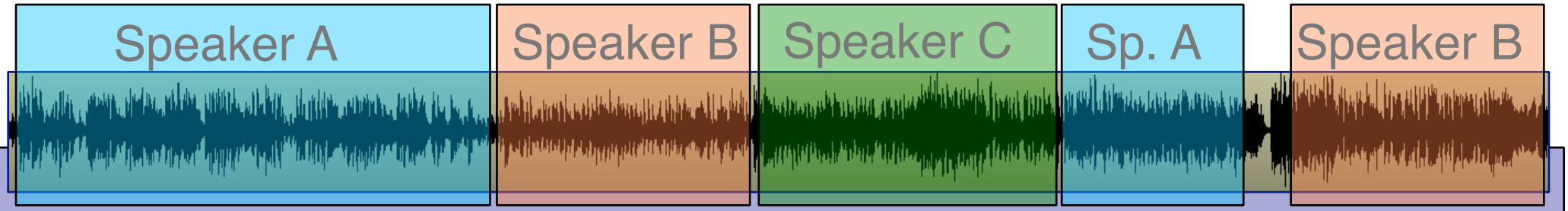
Speaker state domination



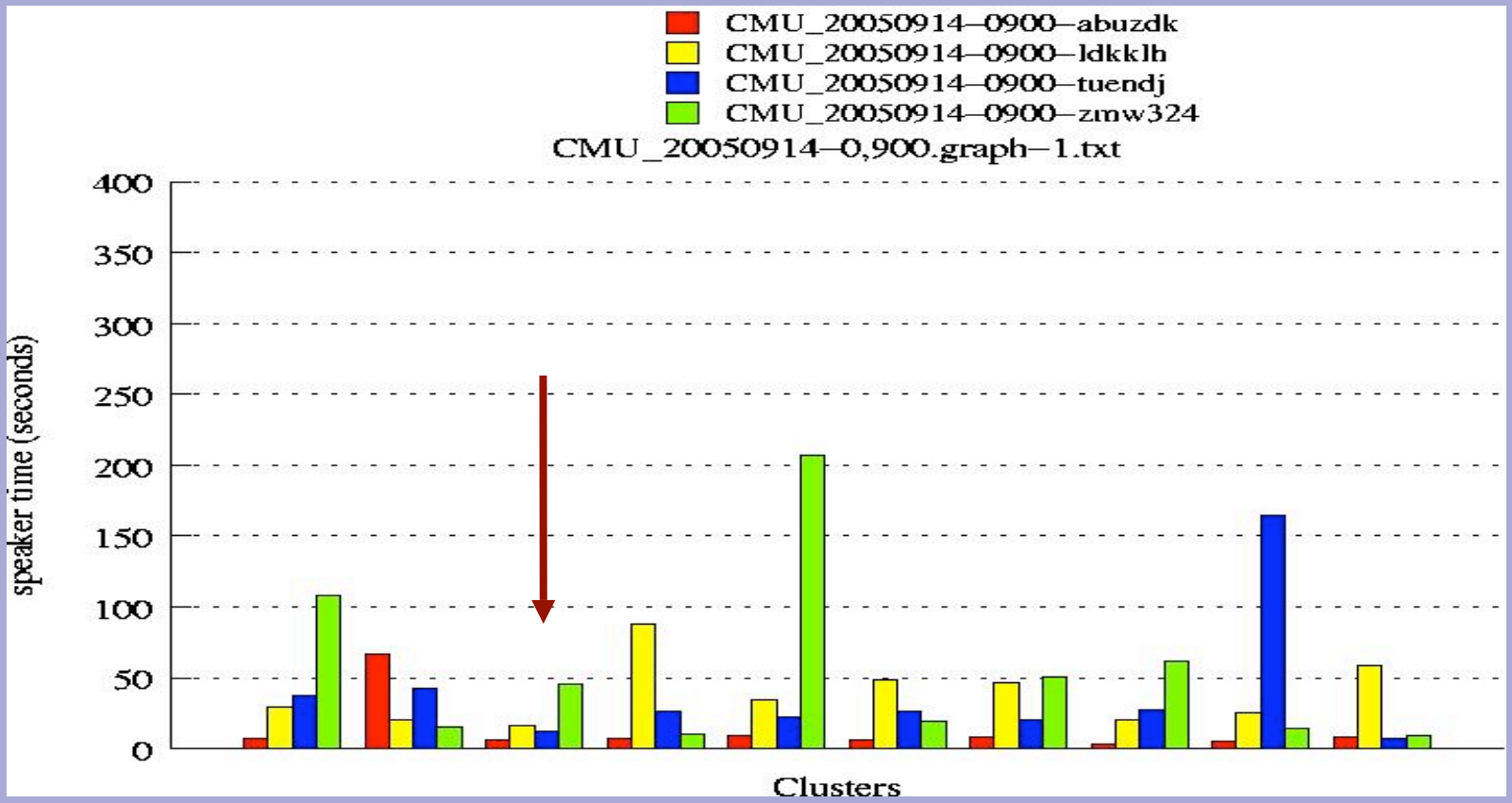
AMI AUGMENTED MULTI-PARTY INTERACTION

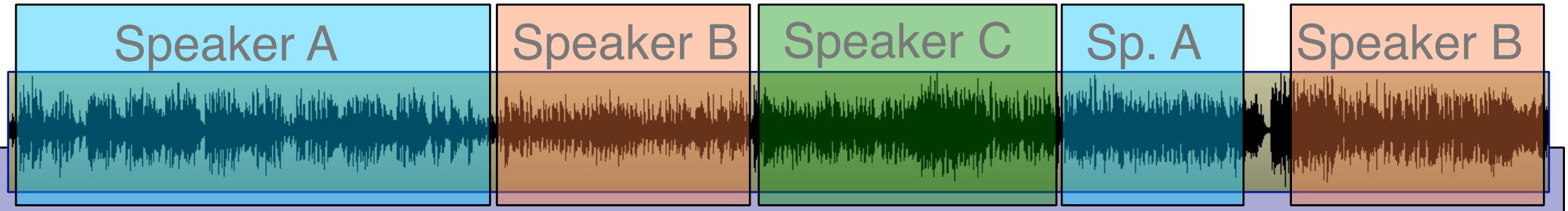
TNO | Knowledge for business

University of Twente
The Netherlands

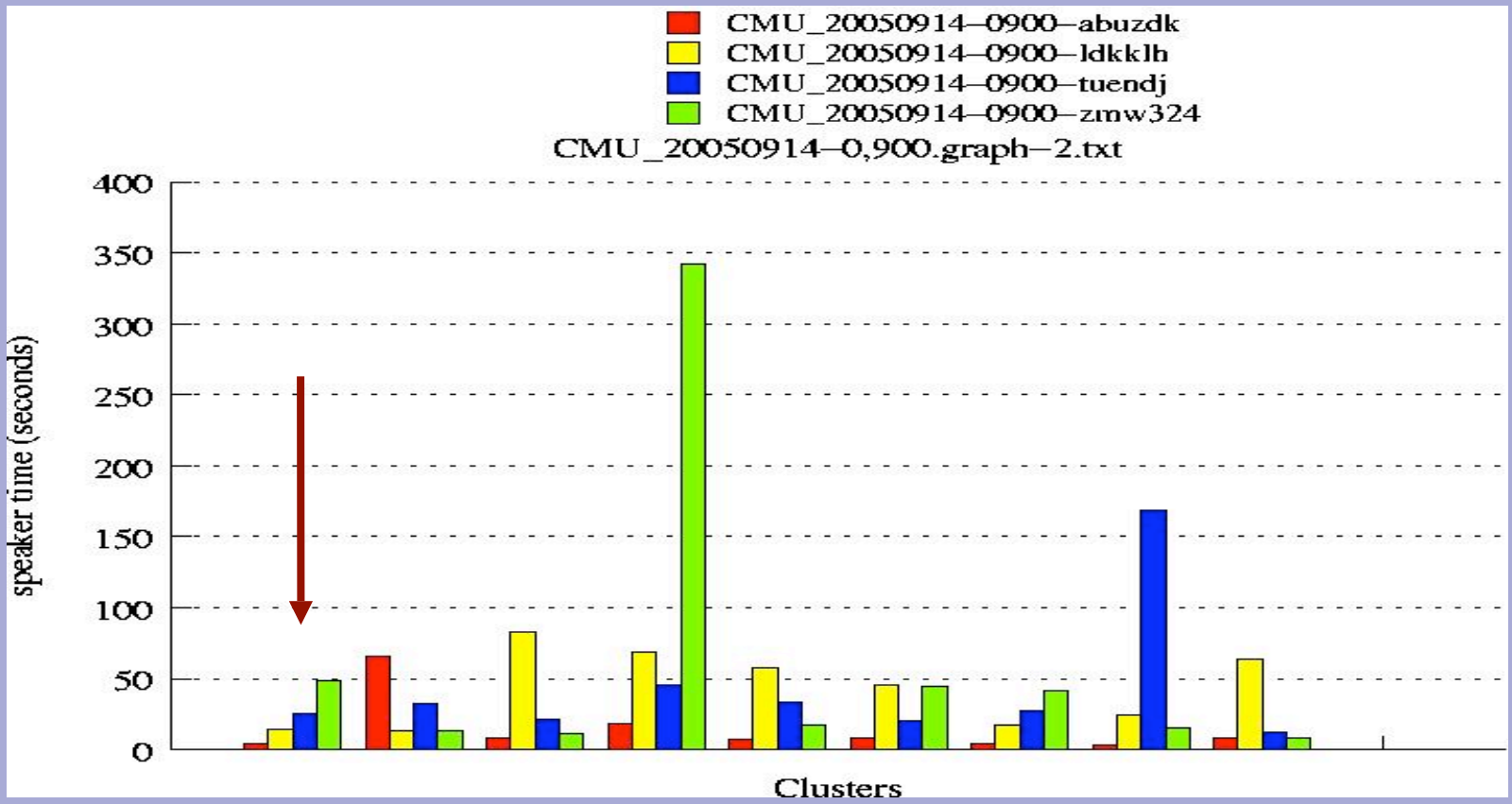


Speaker state domination





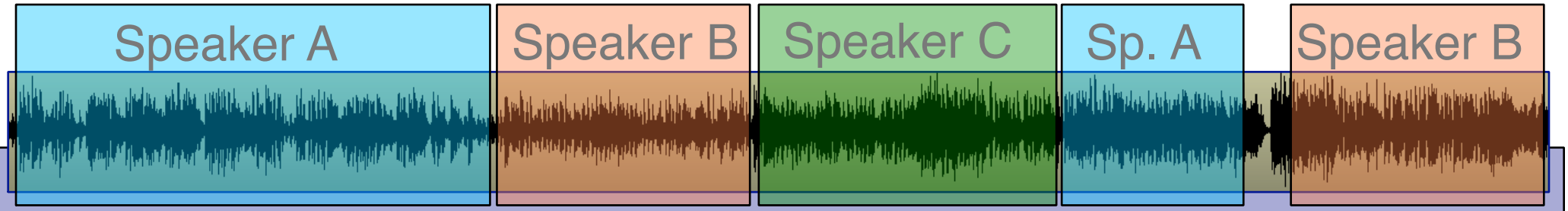
Speaker state domination



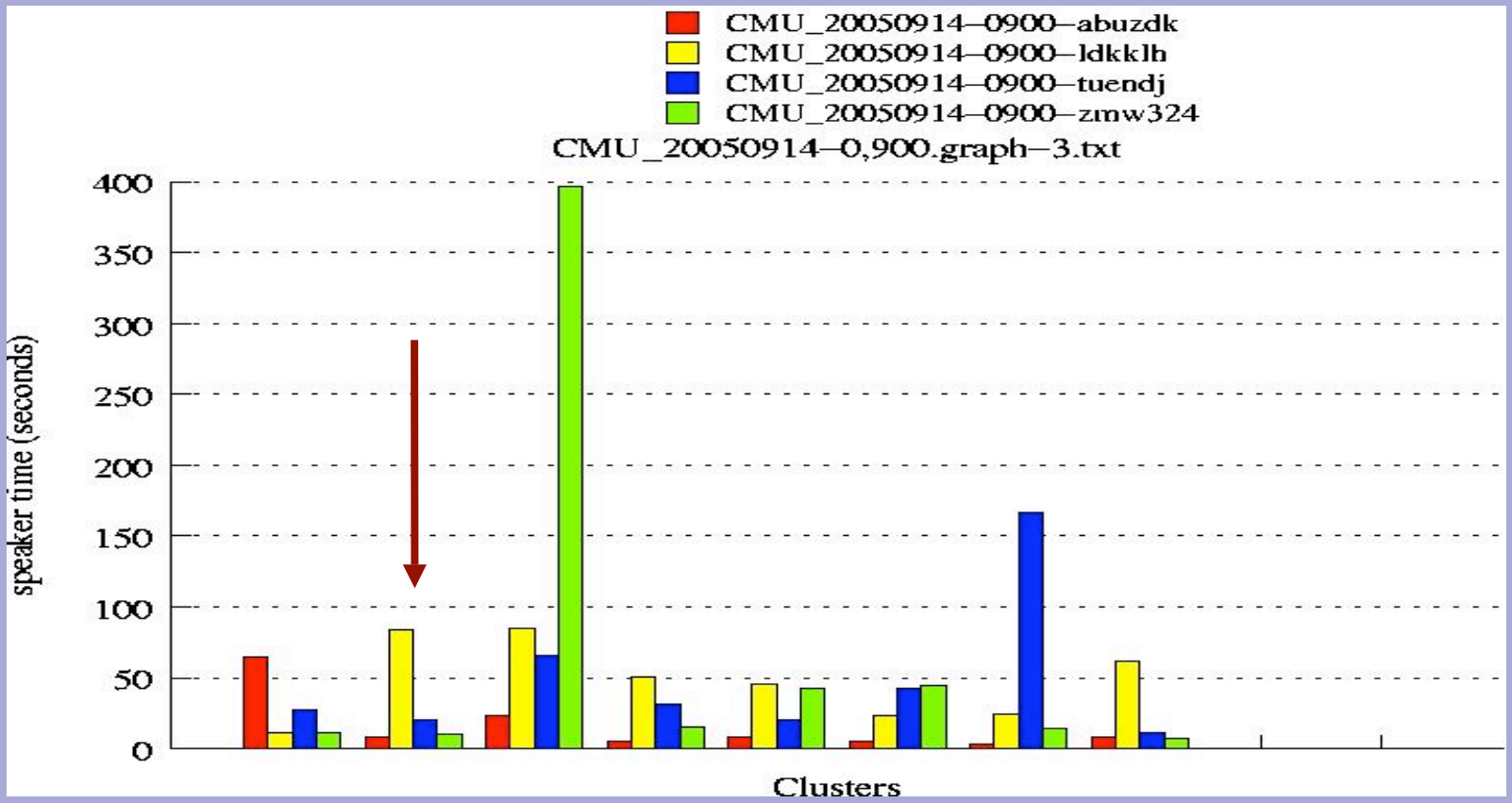
AMI AUGMENTED MULTI-PARTY INTERACTION

TNO | Knowledge for business

University of Twente
The Netherlands



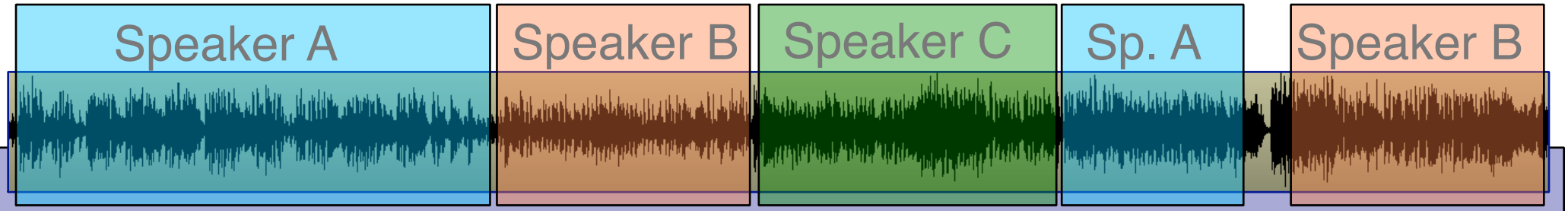
Speaker state domination



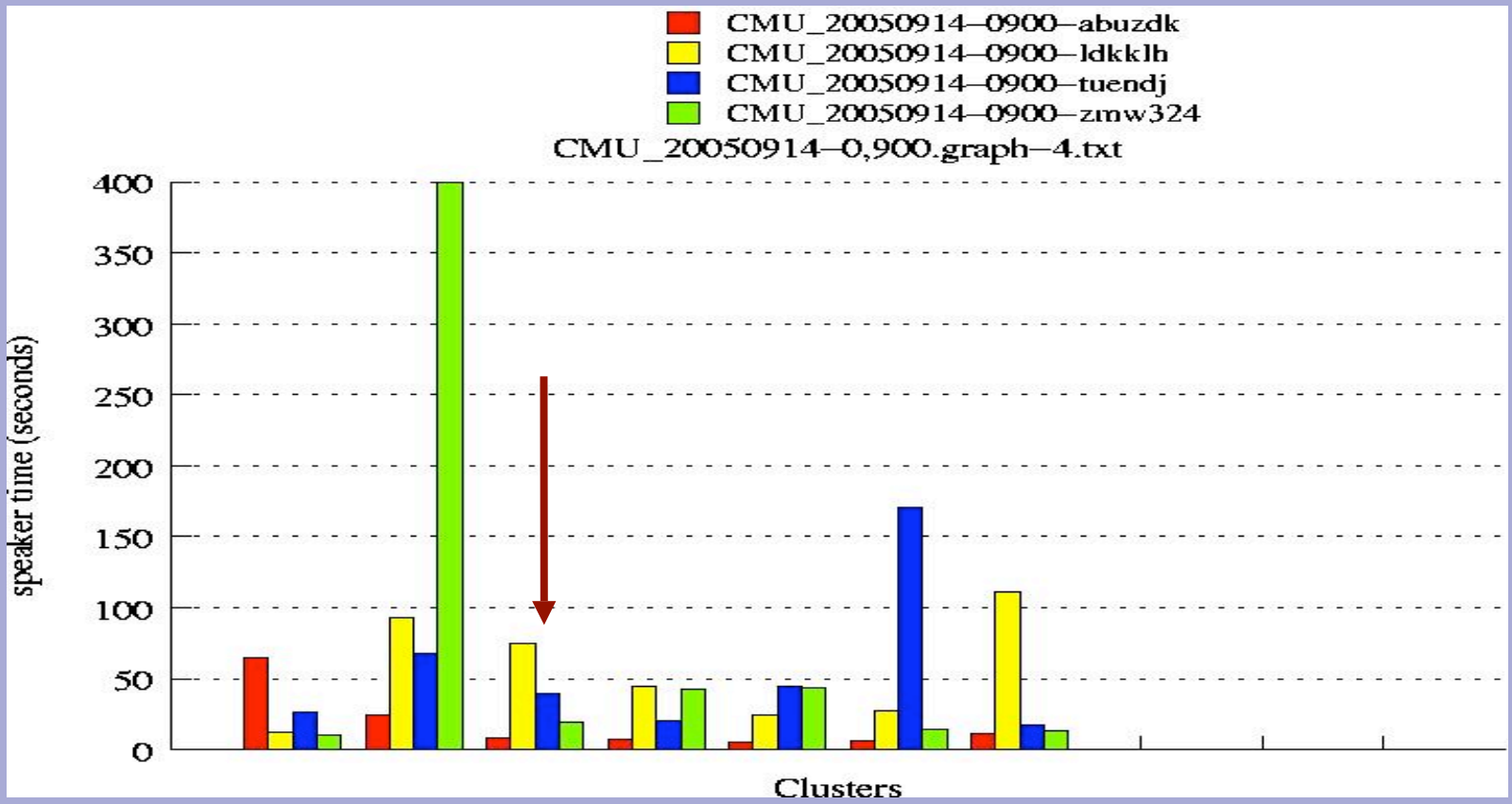
AMI AUGMENTED MULTI-PARTY INTERACTION

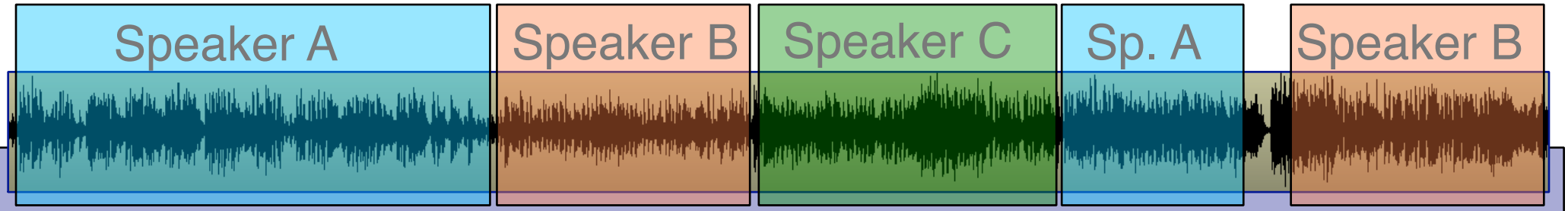
TNO | Knowledge for business

University of Twente
The Netherlands

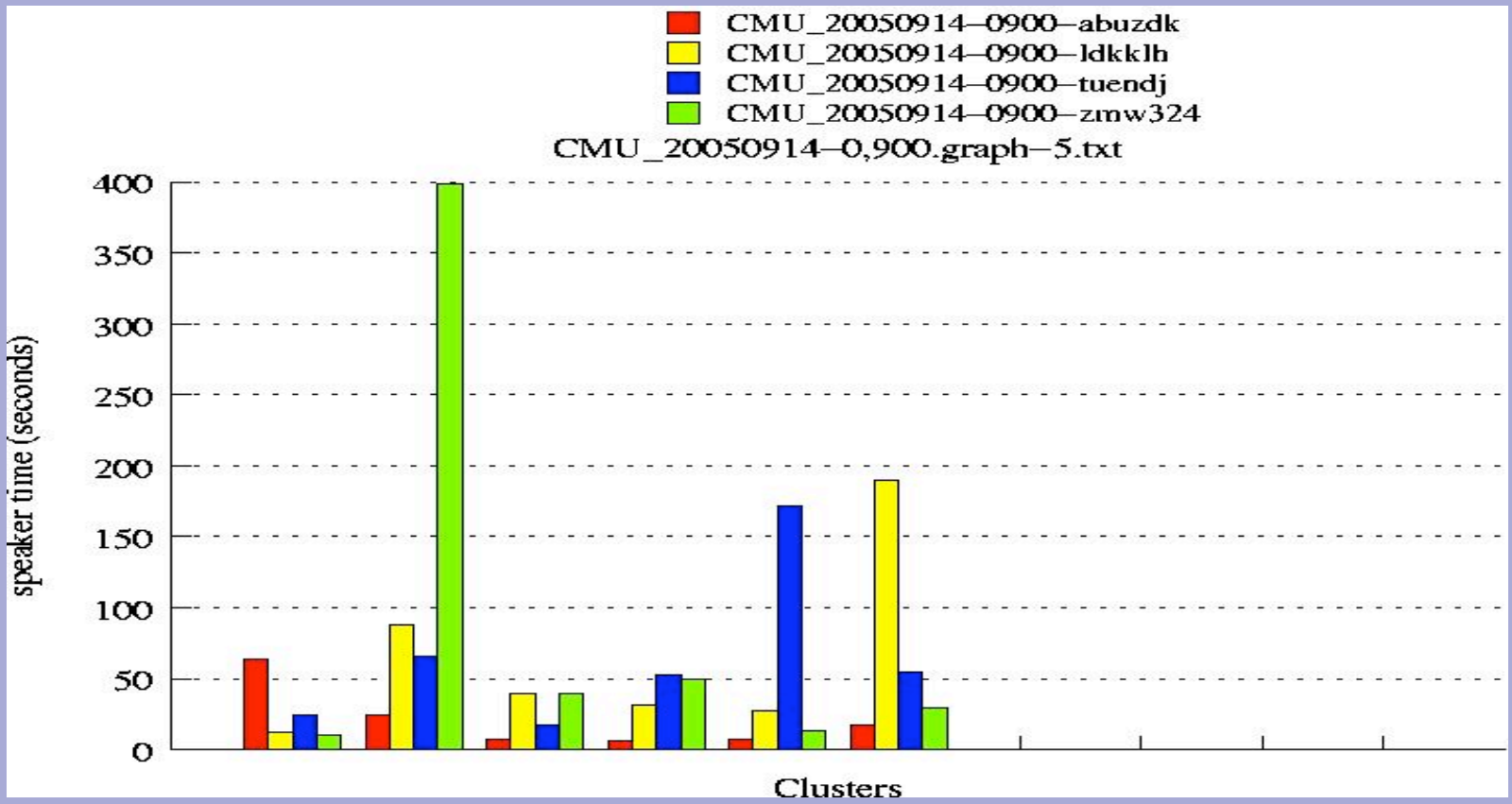


Speaker state domination





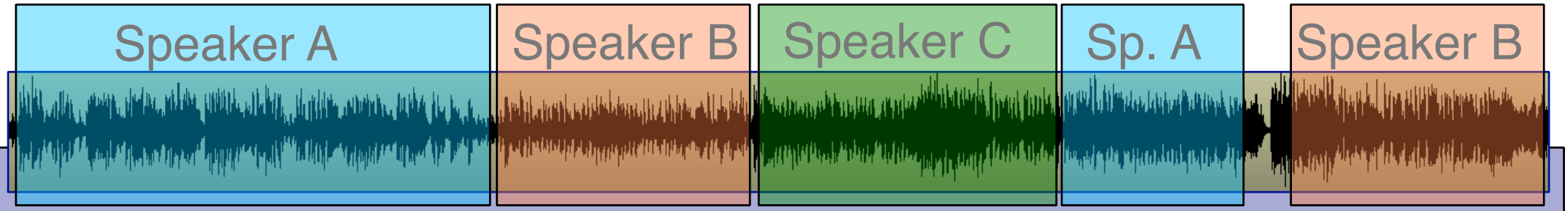
Speaker state domination



AMI AUGMENTED MULTI-PARTY INTERACTION

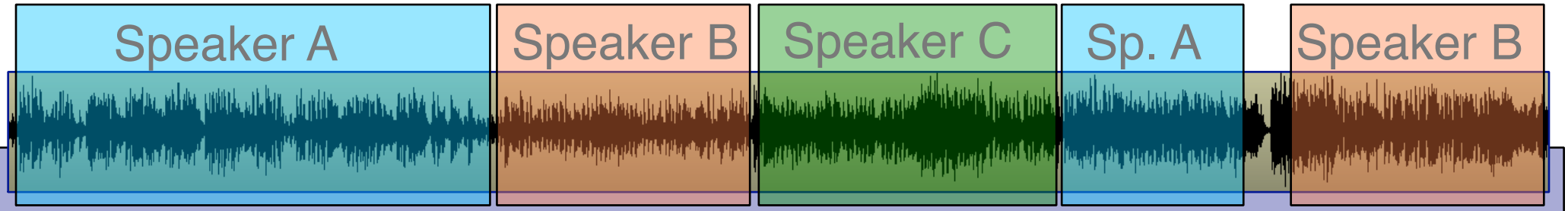
TNO | Knowledge for business

University of Twente
The Netherlands



Future work

- Refined stop criterion for Cut&Mix
- Develop method to make initial states more dominated by single speakers
- Other approach for the initial number of states



Questions?